The Dissertation Committee for Peter Toth
certifies that this is the approved version of the following dissertation:

Essays In Econometrics

Committee:

Jason Abrevaya, Supervisor

Sukjin Han

Haiqing Xu

Shakeeb Khan

Essays In Econometrics

by

Peter Toth,

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2018

Dedicated to my family and friends.

Acknowledgments

I wish to thank my Committee Members and the Econometrics Group at UT Austin. I got great support from my mentoring professors such as Valerie Bencivenga, Eugenio Miravete and Maxwell Stinchcombe. I also thank Austin Bean, Carlos Herrera, Daria Pus and Xinchen Gu for their continuing support and insightful comments in the last five years.

Essays In Econometrics

Publication No. _____

Peter Toth, Ph.D.
The University of Texas at Austin, 2018

Supervisor: Jason Abrevaya

The first chapter of this dissertation considers a semiparametric version of the network formation model of Graham (2017). The two-way fixed-effects binary choice model allows for homophily and degree heterogeneity, but unlike Graham (2017) leaves the distribution of pair-specific unobservables unspecified. Identification of the slope parameters and fixed effects follows from a novel approach that does not rely on distributional assumptions. The identification strategy suggests an estimator for the slope parameters based upon tetrads of nodes within the network. A computationally simple version of this estimator is shown to be consistent with a non-parametric convergence rate. A consistent estimator of the fixed effects is also provided.

The second chapter discusses the non-parametric extension of the network formation model, when the researcher does not assume the functional form of the distance function. An intuitive way for the non-parametric extension is to use the parametric estimator for linear indices coupled with a series expansion.

While the technique is generally applicable, it comes with the caveat that the identification of the models must be assured a priori. After demonstrating the applicability of the method on classical models of Manski (1987) and Han (1987), we prove the nonparametric identification of the distance function for the network formation model, and define the corresponding series estimator. We give a proof for consistency, and also analyze the rate of convergence.

The third chapter examines the empirical content of the assumption that in a complete information game agents play pure strategy Nash-equilibrium. In particular, we focus on the identification of the strategic interaction effects as defined in Tamer (2003). We find that the Nash-equilibrium assumption restricts the joint density of the unobservables in a way that allows us to connect the underlying identification problem to photo stitching, a well-known question in computer science. In the view of this intuition, some of the earlier results in the literature are reinterpreted, and the main proposition shows how the framework can be used to find sufficient assumptions for identification without specifying the distribution of unobservables.

Table of Contents

Acknowledgments iv

Abstract v

List of Tables viii

List of Figures ix

Chapter 1. Semiparametric estimation in network formation models with homophily and degree heterogeneity 1

1.1 Introduction . . . 1
1.2 Model . . . 6
1.3 Identification arguments . . . 9
1.3.1 Notes on notation . . . 9
1.3.2 Identification of the coefficients on homophily terms . . . 10
1.3.2.1 Screening lemma . . . 13
1.3.2.2 Identification lemma . . . 19
1.3.2.3 Tetrad inequality identification . . . 25
1.3.3 Identification of fixed effects in the linear index model . . . 27
1.4 Estimation . . . 31
1.4.1 Tetrad inequality estimator . . . 31
1.4.1.1 Consistency of the tetrad inequality estimator . . . 32
1.4.2 The simplified inequality estimator . . . 36
1.4.2.1 Asymptotic theory of the simplified tetrad inequality estimator . . . 37
1.4.3 Estimator for the fixed effect-differences . . . 42
1.5 Some extensions and additional topics . . . 44
1.5.1 Partial identification . . . 45
1.5.2 The $\beta = 0$ case . . . 47

1.5.3 Identification of F . . . 49
1.5.4 Information about the distance function . . . 49
1.6 Monte Carlo simulation . . . 51
1.7 Conclusion . . . 56
1.8 References . . . 57
1.9 Appendix . . . 62
1.9.1 Preliminary claims . . . 62
1.9.2 Screening lemma . . . 63
1.9.3 Tetrad inequality identification . . . 65
1.9.4 Fixed effects identification . . . 73
1.9.5 Consistency of the infeasible estimator . . . 74
1.9.6 Consistency of the feasible estimator . . . 75
1.9.6.1 Effect of the first stage . . . 75
1.9.6.2 Consistency proposition . . . 76
1.9.7 Convergence rate of the simplified estimator . . . 77
1.9.7.1 Rate of mistake probability . . . 77
1.9.7.2 Hoeffding-decomposition . . . 81
1.9.7.3 Applying the HPS-lemma . . . 85
1.9.8 Estimation of fixed effects . . . 86

Chapter 2. Nonparametric identification of distance functions in network formation models with fixed effects 93
2.1 Introduction . . . 93
2.2 Simplified problem and main idea . . . 97
2.2.1 Main idea . . . 101
2.2.2 Application and identification caveat . . . 104
2.3 Nonparametric estimation of distance functions . . . 108
2.3.1 Identification . . . 113
2.3.1.1 Information from the zero-property . . . 113
2.3.1.2 Using the Taylor-expansion . . . 118
2.3.1.3 Alternative assumptions . . . 120
2.3.2 Estimation . . . 121
2.4 Monte Carlo simulation . . . 123

2.5 Conclusion . . . 125
2.6 References . . . 126
2.7 Appendix . . . 131
2.7.1 Proof of Lemma 9 . . . 131
2.7.2 Proof of Lemma 10 and Proposition 4 . . . 132
2.7.3 Proof of Lemma 12-13 . . . 132
2.7.4 Proof of Proposition 6 . . . 133

Chapter 3. The empirical content of the Nash-equilibrium assumption in discrete games 145
3.1 Introduction . . . 145
3.2 Econometric model . . . 150
3.2.1 Predictions of the economic model . . . 150
3.2.2 Econometric assumptions . . . 151
3.3 Identification argument . . . 154
3.3.1 The joint probability distribution function and the Nash assumption . . . 155
3.3.2 Identification of the strategic interaction vector . . . 156
3.3.2.1 Photo stitching and identification . . . 158
3.4 Simple examples for sufficient assumptions . . . 166
3.4.1 Restrictions on peaks . . . 167
3.4.2 Restriction on the slope of the isodensity and the gradient . . . 169
3.5 Conclusion . . . 172
3.6 References . . . 174
3.7 Appendix A . . . 177
3.7.1 Proof of Proposition 7 . . . 177
3.7.2 Proof of Corollary 3 . . . 182
3.7.3 Proof of Proposition 8 . . . 183
3.7.4 Proof of Proposition 9 . . . 184
3.8 Appendix B . . . 186
3.8.1 Corresponding estimator of Δ . . . 186
3.8.2 Identification of the slope parameters . . . 188

Bibliography 191

List of Tables

1.1 Simulation results for Graham's tetrad logit and the tetrad inequality estimator. 53
1.2 Monte Carlo simulation results for the simplified estimator. 55

2.1 Monte Carlo results for the conditional series estimator. 124

3.1 Pure strategy Nash-equilibria of the entry game. 150

List of Figures

1.1 Stylized picture of the data set for $N = 4$. 2
1.2 Screening with $p = 2$. 15
1.3 Screening with $p = 2$ using a special screening set. 16
1.4 Screening values switching signs with $p = 2$. 17
1.5 Information content of the double-differences of the observables if we observe that the screening values switch signs. 21
1.6 Sufficient variation to differentiate between b and β for $p = 3$. . . . 24
1.7 FE identification with $p = 2$. 29
1.8 Definition of $X_{ij}^* = \{x \in \mathbb{R}^p : S_{ij}(x) = 0\}$. 50

3.1 Notation summarized. 156
3.2 A photo stitching problem and its solution. 160
3.3 Notation summarized. 163
3.4 Notation summarized for the proofs. 178

Chapter 1

Semiparametric estimation in network formation models with homophily and degree heterogeneity

1.1 Introduction

Networks are important for many economic problems. For instance, they are present in the theory of production (supplier networks), international trade, and job search, and they also serve as the mechanism for peer effects in education economics and other fields. The typical dataset for a network would consist of nodes (economic agents), observable characteristics of the economic agents, and indicator variables denoting whether links exist between any pair of nodes. As an example, the network depicted in Figure 1.1 is a friendship network among high-school students with some observed characteristics, where a link between nodes indicates a friendship.

Following Graham (2017), we focus on a network formation model that is characterized by

$$D_{ij} = \mathbb{1}[W_{ij}\beta + A_i + A_j \geq U_{ij}], \tag{1.1}$$

where D_{ij} is equal to one if a link exists between nodes i and j and zero otherwise. W_{ij} is a vector of observed distances between the nodes that can be written as a known vector-valued function of the observables. The $W_{ij}\beta$ term allows for

homophily, the phenomenon that nodes with similar observables connect with a higher probability. A_i and A_j are unobserved fixed effects that may be related to observables. In particular, the joint distribution of observables and fixed effects is unrestricted. These fixed effects allow the network to exhibit *degree heterogeneity*, whereby some nodes have many (or few) connections for some reason that is not completely explained by observables. Finally, U_{ij} is a pair-specific unobservable. Whereas Graham (2017) and other papers (Dzemski (2016), Jochmans (2017)) consider the case where U_{ij} is parametrically specified, the focus of this paper is the case where the distribution of U_{ij} is left unspecified.

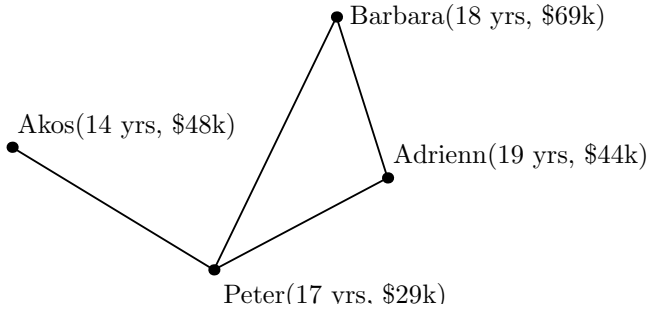


Figure 1.1: Stylized picture of the data set for $N = 4$. The observables are age and family income. The older nodes are all connected, while the younger node (Akos) is only connected to Peter.

This paper provides a semiparametric identification strategy that does not rely on distributional assumptions on the unobservables U_{ij}. In particular, subject to necessary normalizations, we provide a novel approach that allows us to identify the relative importance of the homophily dimension (β) and the individual fixed effects (A_i). Our results are based on a "screening lemma," which formalizes the intuition that if node i has a higher probability of linking to a set of nodes (screening

set) with fixed observable characteristics than node j, then i is either closer to this set or has higher fixed effect than j. For another screening set (with different values for the fixed characteristics vector) for which node j has higher linking probability than node i, the fact that the ordering of fixed effects remains the same indicates that the change in the probability comparison was driven by the change in the homophily terms. These comparisons of screening sets lead to an implication about the sign of the double-difference of the observable linear index that leads to identification of β up-to-scale. With β identified, the identification of the fixed effects A_i also follows from the screening lemma. If we find an observable vector such that the link probabilities of i and j with the associated screening set are equal, then the differences in the observable linear index must exactly offset the difference in the unobserved fixed effects. This identification framework can be applied to a more general class of models to achieve the identification of the distance function and the fixed effects to the same normalization, which is the topic of Toth (2018).

Based on the identification strategy, we propose a *tetrad inequality estimator* for the slope parameters. The proposed estimator is computationally expensive, as it involves the enumeration of $\binom{n}{4}$ tetrads, where n is the number of nodes. For this reason, we also consider a simplified version of the estimator that requires only $\binom{n}{2}$ pairwise computations in its objective function. Asymptotic theory is developed for the simplified version of the estimator. The estimator is consistent with a non-parametric convergence rate. Following the identification strategy, an estimator of the individual fixed effects A_i is also proposed, and its asymptotic

properties are analyzed.

Our semiparametric approach avoids the possible misspecification associated with estimators based upon parametric assumptions on U_{ij}. Moreover, we demonstrate that the identification of the network formation model does not hinge critically upon the parametric specifications considered previously in the literature. To our knowledge, this paper is also the first to provide an identification and estimation strategy for the fixed effects values without a parametric assumption on the error disturbances.

There is a vast literature on networks, and therefore we focus on the parts of that literature most related to our current paper. The information content of the average degree and the degree sequence has been considered by Barabasi and Albert (2002), Lovasz (2012), and others. De Paula, Shubik, and Tamer (2016) provide identification and estimation results for a dynamic network formation model. Leung (2016a,b) uses covariances and averages of links for statistical inference for identification in the presence of homophily and externalities. Another strand of the literature, including Mele (2015), Sheng (2014), and Menzel (2015), focuses on using the relative frequencies of certain subgraphs in order to estimate the parameters of the respective structural model. Unrestricted degree heterogeneity interferes with the identification of externalities caused by social interactions in the case when the researcher only observes one large network, and including both mechanisms in the model would require multiple observations of the same network (see Graham (2016)). For this reason, the model considered in this paper should be viewed as complementary to the models with externalities. As for the identification and

estimation of fixed effect parameters, a recent paper from Jochmans and Weidner (2017) solves the problem in the case when homophily effects are missing from the network.

Our network formation model is also closely related to a two-way fixed effect linear panel data model for binary choice when the number of time periods and individuals are both high. While there are results for this model using a likelihood approach after appropriate parametrization of the unobservables (Chernozhukov et al 2015), the semipametric version of the model has not been considered in the literature. The approach taken here can be viewed as a two-dimensional generalization of Manski (1987), which provides an estimator for the semiparametric binary choice model with (one-way) individual fixed effects. The presence of two fixed effects in the model significantly complicates matters, relative to Manski (1987), since simple first-difference comparisons are not sufficient.

The outline of the paper is as follows. Section 2 describes the semiparametric network formation model in detail. Section 3 provides the main identification arguments, including the screening lemma described above. Section 4 defines the tetrad inequality estimator for β, as well as its simplified version, and an estimator for the fixed effects A_i. Asymptotic properties of these three estimators are established. Section 5 provides some additional results and extensions of interest. Specifically, we show that the model gives information about the overall presence of homophily effects ($\beta \neq 0$), discuss the identification of the cumulative distribution function (cdf) of unobservables, and provide simple extensions to the model where the identification strategy still applies. Section 6 provides Monte Carlo simulation

evidence on the performance of the proposed estimators, and Section 7 concludes. Proofs are provided in the Appendix.

1.2 Model

In this section we introduce the model and discuss its main limitations. Throughout the paper we assume that the researcher observes N nodes in the network, and that each node i has a row vector of observable characteristics X_i, which has p elements. If we vertically stack the X_i vectors, we get the X matrix. The researcher also observes the $N \times N$ adjacency matrix D, which encodes in its (i,j)th element whether the ith and jth node is connected ($D_{ij} = 1$), or not ($D_{ij} = 0$). We assume that N is large, but the network is only observed once.

Every node i is equipped with an unobservable characteristic A_i (fixed effect), and every i, j pair (dyad) has an unobservable dyad-level characteristic U_{ij}. Let A be the $N \times 1$ column vector the ith element of which is A_i. The data generating process is modeled by the link formation equation described in the following assumption.

Assumption 1 (Link formation.)

The vector of observable and unobservable individual characteristics, (X_i, A_i) *are independently and identically distributed through the nodes. The links in the networks are formed according to*

$$D_{ij} = \mathbb{1}[W_{ij}'\beta + A_i + A_j \geq U_{ij}], \tag{1.2}$$

where

$$W_{ij} = w(X_i, X_j) = \begin{pmatrix} w_1(X_i^1, X_j^1) \\ w_2(X_i^2, X_j^2) \\ ... \\ w_p(X_i^p, X_j^p) \end{pmatrix}, \tag{1.3}$$

for some known functions $w_k : \mathbb{R}^2 \to \mathbb{R}, k \in \{1, 2, ..., p\}$.

In our motivation these w_k functions are distances, like

$$w_k(x, y) = |x - y|,$$

or

$$w_k(x, y) = (x - y)^2,$$

so we will assume some of the metric properties.

Assumption 2 (Distance function.)

We require that for every k*, the* $w_k(.,.)$ *function is symmetric and continuous in its two arguments, and*

$$\forall (x, y) \in \mathbb{R}^2 : x = y \Leftrightarrow w_k(x, y) = 0. \tag{1.4}$$

To simplify the analysis, we assume that the network is symmetric.[1] For example, if we model a network of friends, the friendships are assumed to be reciprocated. Symmetry corresponds to the economic assumption that the agents can share the gain from establishing a connection. This assumption is embodied in the condition that the index in the link formation equation is symmetric in i and j.

[1] The links are undirected.

The linearity in Assumption 1 makes the coefficients on the distances constant with respect to the distance characteristics *and* the fixed effects. We frequently call $W_{ij}'\beta$ the 'observable index', and $A_i + A_j$ the unobservable part of the index. Our fully linear specification restricts the trade-off between the observable and unobservable parts of the index. The restriction on the trade-offs between the unobserved characteristics becomes important to give meaning to the fixed effects and to identify them up to scale and location normalization. For the arguments in this paper the linearity of the observable index is important, but it turns out that it can be weakened for identification purposes. However, that would require a different use of screening, and so it is not the topic of another paper.

Perhaps the most important limitation of our model is that we abstract from social interactions. This encompasses two assumptions. First, by Assumption 1 the index in the link formation equation between two nodes is only influenced by their observables and fixed effects. This helps to rule out that a third node's characteristics or links have an effect on the formation of the link. Second, we need to assume that the pair-specific unobservables (U_{ij}) are exogenous with respect to the whole X matrix and A vector, and also independent of each other.

Assumption 3 (Exogeneity, iid U_{ij}-s.)
$U_{ij} \perp (X, A)$ *for all* i, j*. Moreover, the* U_{ij} *random variables are independent of each other, and identically distributed with a cumulative distribution function* F*.*

Assumption 3 is central to our results, and together with Assumption 1 it rules out mechanisms that involve any third node's characteristics or outcomes in

the link formation equation. This allows us to abstract from problems including multiple equilibria, and focus on the question at hand.

1.3 Identification arguments

In this section we present the identification argument. Our goal is to identify the coefficients on the elements of the W_{ij} vector and the fixed effect values up to appropriate normalizations.

1.3.1 Notes on notation

In the following we will use vectors frequently. For a vector z, we denote the kth element of this vector by z^k. Moreover, $||z||$ denotes the Euclidean-norm of the vector.

Given $\epsilon > 0$, and $x \in \mathbb{R}^l$, let us define $B_\epsilon(x)$ as the open ϵ-ball around x:

$$B_\epsilon(x) = \{z \in \mathbb{R}^l : ||x - z|| < \epsilon\}.$$

We denote the support of a random variable Z as $Supp(Z)$, and for our purposes it is defined as

$$Supp(Z) = \{z \in \mathbb{R} : \forall \epsilon > 0, P(Z \in B_\epsilon(z)) > 0\}.$$

The support of a finite random p-vector X is defined analogously. The support of a random variable Z, conditional on another random variable S taking the value s is denoted as

$$Supp(Z|S = s) = \{z \in \mathbb{R} : \forall \epsilon > 0, P(Z \in B_\epsilon(z)|S = s) > 0\}.$$

1.3.2 Identification of the coefficients on homophily terms

First we look at the identification result concerning the coefficients in the linear index. After describing the necessary normalization of the parameters of interest, we discuss the identifying assumptions related to the distribution of observables and the fixed effects. Just as in Manski (1985), besides the independence assumption, we need sufficient variation of the observables for identification. However, since we have endogenous unobservables (the fixed effects) in this model, we need to make some regulatory assumptions on their support as well. After stating these remaining conditions, we will see that the coefficient vector is identified up to scale. We close this subsection by defining an identifying statistic in the sense of Pakes and Pollard (1989).

As a first step of our analysis, we need to apply the scale normalization for β by assuming that

Assumption 4 (Scale normalization.)

$$||\beta|| = 1.$$

With this latest assumption the β is a direction on the unit sphere in $\mathbb{R}^p$.[2] This normalization is necessary, as the scale parameter of F is not restricted, and the assumption is in line with the literature of binary outcome models (see for example Manski 1988). With Assumption 4 we implicitly assume that at least

[2]Note that this is a compact set in $\mathbb{R}^p$ under the Euclidean norm.

one of the homophily dimensions have an effect on the probability of two nodes connecting. For the discussion of the case when $\beta = 0$, please refer to Section 5 of this paper.

While Assumption 1-3 are the main identifying assumptions that provide information about the parameters, we need further support assumptions to ensure point identification.

Assumption 5 (Overlapping conditional support of A_i-s)

A_i is distributed continuously on the same support, conditional on $x = X_i$ for any $x \in Sup(X_i)$.

$$Supp(A_i|X_i = x) = Supp(A_i),\ \forall x \in Supp(X_i).$$

Remark 1. {The condition above is a sufficient assumption, and it can be substantially weakened. For example, is enough to require that there exists a point $C \in \mathbb{R}$ and an $\epsilon > 0$, such that

$$\inf_{x \in Supp(X_i)} P[A_i \in B_\epsilon(C)|X_i = x] > 0$$

Even this condition can be weakened, but some form of the common support assumption is important for the arguments in the proofs to work. Another type of identifying assumption would be to assume that the conditional support of fixed-effect differences includes the support of the differences in observable distances. We consider those type of assumptions more restrictive for practical purposes.}

As we will see below, the differences in the observable distance vectors are going to identify the (direction of) β. For this reason, given X_i and X_j (i and

j being nodes on the network) let us introduce the row-vector $\Delta_{ij}(x)$, the kth element[3] of which is

$$\Delta_{ij}^k(x) = w_k(X_i^k, x^k) - w_k(X_j^k, x^k),$$

for $k = 1, 2, ..., p$. Intuitively, just like in the seminal article of Han (1987), if we would like to point-identify β, we need that $\Delta_{ij}(X_k)$ takes every directions on $\mathbb{R}^p$. This is ensured by the following assumption.

Assumption 6 (Support assumption for X_i.)

X_i is distributed continuously, and its support is not a subset of a proper subspace in $\mathbb{R}^p$.

Assumption 6 is a strong support assumption, as it does not allow for discrete observables.

Remark 2. {There are alternative assumptions that are not as restrictive as the previous one above. Following the same arguments as for example in Kline (2015), analogous results to Proposition 1 can be extracted if at least one of the observables is continuously distributed with a large enough support. In general, there is a trade-off between requiring large support and allowing for discrete observables. Here we chose to require every observable to be continuously distributed, and make no large support assumption. The other extreme case would be to require unbounded support in every element of the X_i vector, but allow for discrete variables in arbitrary number of dimensions. As this issue is not the focus

[3] $\Delta_{ij}(x)$ has length p

of this paper, and the problem has been treated elsewhere the same way we would handle it here, we continue with Assumption 6.}

1.3.2.1 Screening lemma

Our identification argument consists of two steps. The first step (screening) is describing how conditional probabilities can be used to infer the order of the realized values of the indices. This is expressed in the screening lemma. Using this information, we can proceed to the second stage (regression) to make an identification argument similar to Han (1987).

In this section we take differences in the conditional expectations of D_{ik}-s and D_{jk}-s, where we condition on the observable characteristics of the kth node and the characteristics of i and j, on X_i, X_j, A_i, A_j. We call this process screening, as intuitively we compare the 'performance' of node i and j using some exogenously selected group of nodes as 'testing ground'. Define

$$h(x_1, x_2, a_1, a_2; x_3) = E[D_{ik} - D_{jk} | X_k = x_3, A_i = a_1, X_i = x_1, A_j = a_2, X_j = x_2]$$
$$\delta_{ij}(x) = h(X_i, X_j, A_i, A_j; x), \tag{1.5}$$

where the expectation is taken over U_{ik}, U_{jk} and A_k, as A_i, A_j and X_i, X_j are conditioned upon. We call $\delta_{ij}(x)$ the screening value. It is a function of the vectors (X_i, A_i), (X_j, A_j) and the constant vector x. The vector x determines the screening set, the nodes in the population with characteristic vector x. Note that given the nodes i and j, the screening value $\delta_{ij}(x)$ is identified. Practically, we can fix the nodes i and j when computing the expectation, so that we guarantee the appropriate fixing of the A_i, A_j and the X_i, X_j, respectively.

Figure 1.2 visualizes the information given by the observables and the screening probabilities when $p = 2$. We can place two nodes i, j and the screening set to the plane spanned by the observables, as they can be represented by $X_i = x_i, X_j = x_j$ and x, respectively. The width of the segments between the points signifying the nodes i, j and the screening set represent the values $P[D_{ik}|X_i = x_i, A_i = a_i, X_k = x_k]$ and $P[D_{jk}|X_j = x_j, A_j = a_j, X_k = x_k]$. The realization of $\delta_{ij}(x)$ is positive, since the segment connecting the blue cross (i) and the black ball (screening set) is thicker than the segment connecting the red cross (j) and the black ball. At this point, we do not know if we observe these relations because the fixed effect of node i is larger than the fixed effect of node j, or because node i is closer to x than node j.

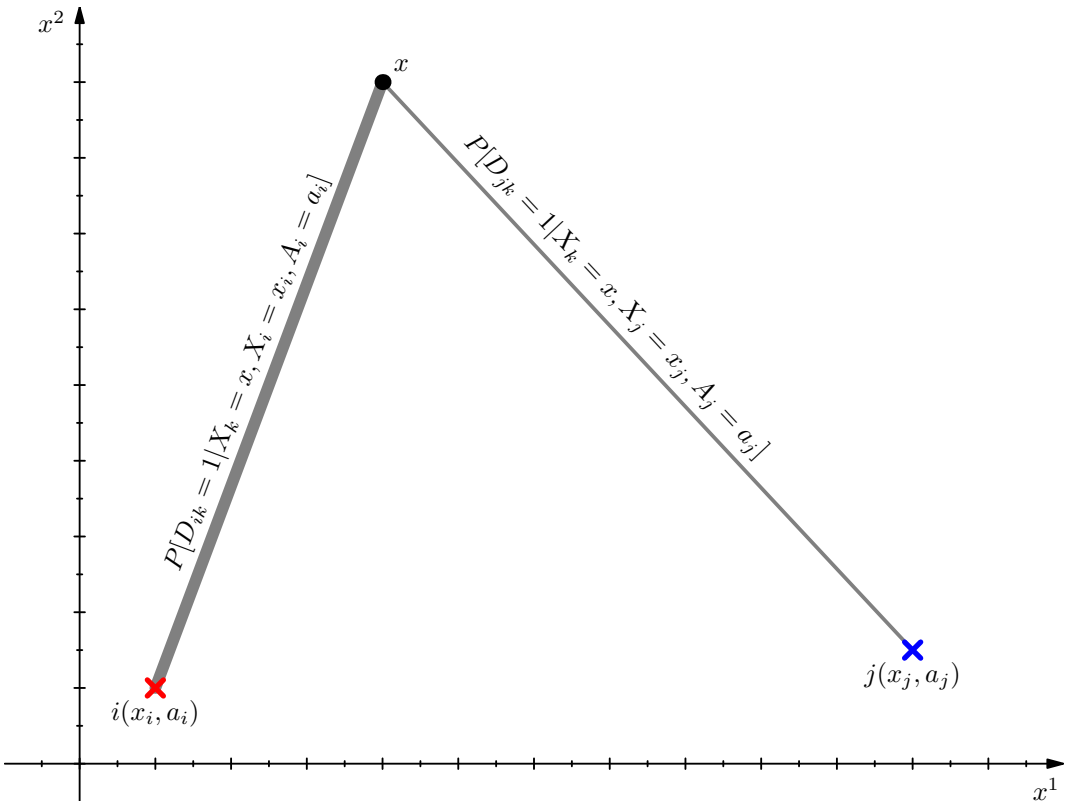


Figure 1.2: Screening with $p = 2$. The figure shows the nodes i, j and the observable values x defining the screening set on the plane of observables. The width of the segments connecting points show the size of the connection probabilities between i (and j) and a random node from the screening set (nodes with observable vector x). We can see that i is closer in position of observables (x) and in terms of probabilities to the screening set as well.

How can we use screening then? A simple way to show that calculating these conditional probabilities is useful can be seen in Figure 1.3. There we chose x, the observables defining the screening set to be exactly in the middle of x_i and x_j. The fact that the screening value is negative tells us that $a_j > a_i$, since the screening set is equidistant from i and j by design, so the only factor that could

influence the connecting probabilities is the difference in the fixed effects.[4]

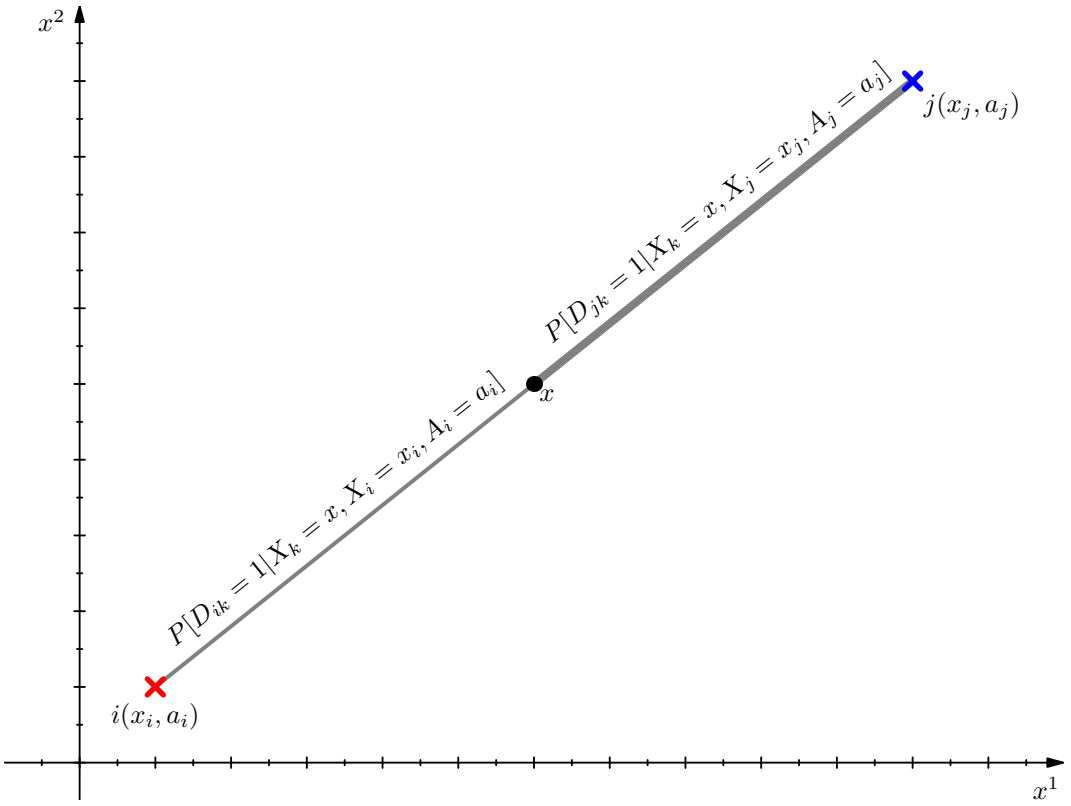


Figure 1.3: Screening with $p = 2$ using a special screening set. The figure shows the nodes i, j and the observable values x defining the screening set on the plane of observables. The nodes in the screening set are equally far away from i and j in both dimensions. The width of the segments tell us that the connecting probability of i is less than that of j with the nodes in the screening set. Given that $W_{ik} = W_{jk}$, this has to reflect the relative sizes of the realized fixed effect values a_i and a_j.

The identification argument in this paper is based on the situation depicted in Figure 1.4. The screening value with respect to the screening set x is positive,

[4]This use of screening is going to be utilized in our forthcoming paper, as the identification strategy corresponding to it does not depend on the linearity of the index.

but it switches sign when we screen i and j with respect to x'. Intuitively, the only thing that could change the sign is the change in the homophily terms, as the fixed effects remained the same. The screening lemma below summarizes what information we use from the screening values.

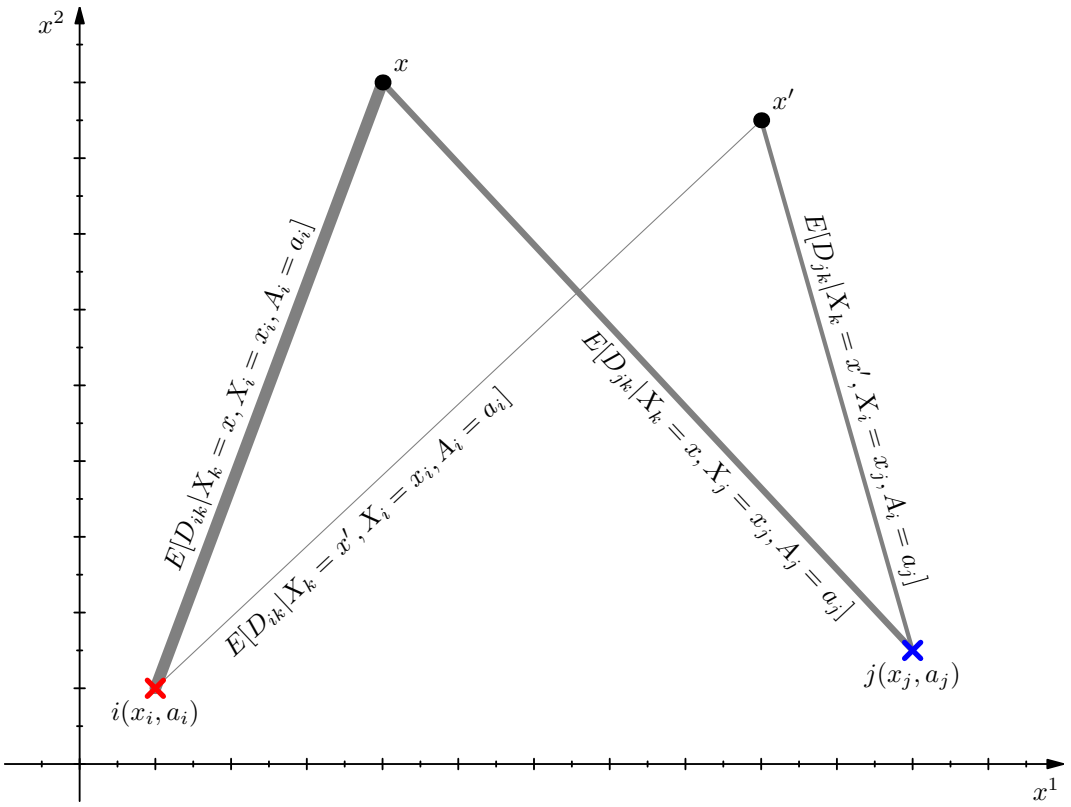


Figure 1.4: Screening values switching signs with $p = 2$. The figure shows the nodes i, j and the observable values x, x' defining the screening sets on the plane of observables. The width of the segments connecting two points show the size of the connection probabilities. Clearly, the realizations $\delta_{ij}(x) > 0$, while $\delta_{ij}(x') < 0$.

Lemma 1 (Screening.)

In the linear index model with Assumption 1-4, for any $x, x' \in Supp(X_i)$

$$\delta_{ij}(x) > 0 \text{ and } \delta_{ij}(x') < 0 \Leftrightarrow \Delta_{ij}(x)\beta > A_j - A_i > \Delta_{ij}(x')\beta.$$

Proof. For the full proof please visit the Appendix. Using Assumption 3,

$$\delta_{ij}(x) > 0 \Leftrightarrow E[F[w(X_i, x)\beta + A_i + A_k] -$$
$$-F[w(X_j, x)\beta + A_j + A_k]|A_i, A_j, X_i, X_j] > 0,$$

where A_k follows the conditional distribution of the unobservable characteristics, when the observable vector takes the value x. We have that regardless the realization a, by the strict monotonicity of F

$$F[w(X_i, x)\beta + A_i + a] > F[w(X_j, x)\beta + A_j + a] \Leftrightarrow \Delta_{ij}(x)\beta + A_i - A_j > 0. \tag{1.6}$$

This means that the integrand in the previous conditional expectation is either exactly always positive, always negative, or always zero, and this only depends on the sign of $\Delta_{ij}(x)\beta + A_i - A_j$. If we integrate (1.6) over a (using the conditional cdf of the unobservables), we get

$$\delta_{ij}(x) > 0 \Leftrightarrow \Delta_{ij}(x)\beta + A_i - A_j > 0. \tag{1.7}$$

The argument works in both directions of the inequality, symmetrically for the case when $\delta_{ij}(x') < 0$, from which the result follows. □

The lemma says that if after changing the screening set (from x to x') the screening value *switches sign*[5], then the change must be driven by the changes in the

[5]This corresponds to the conditional connection probabilities *switching order*.

observable homophily terms. This is true simply because all the other candidates that could change the probabilities either cancel out (A_k-s), or stay fixed (A_i, A_j) while screening.[6]

1.3.2.2 Identification lemma

Define the event G_{ij} as

$$\delta_{ij}(x) > 0 > \delta_{ij}(x')$$

for the (X_i, A_i) and (X_j, A_j) random vectors and $x, x' \in Int\left(Supp(X_i)\right)$, where $Int(A)$ denotes the interior of the set A. If this event happens, we know that

$$[\Delta_{ij}(x) - \Delta_{ij}(x')]\beta > 0 \tag{1.8}$$

by the screening lemma. This inequality means that the true β is in a positive half-space of $\mathbb{R}^p$ defined by $\Delta_{ij}(x) - \Delta_{ij}(x')$ as its normal vector. That is, if G_{ij} happens, we can rule out a half-space where β cannot lie. This is what we can see in Figure 1.5 at stage 1 for $p = 2$. The red semicircle represents the region for the normalized β vector that we could not rule out after observing the $\Delta_{i^1 j^1}(x) - \Delta_{i^1 j^1}(x')$ vector together with the $G_{i^1 j^1}$ event. Then if for some other i^2, j^2 pair in the network the

[6]Unless we specify F, we cannot use more than the order of the probabilities, as the (possibly varying) curvature in F influences the numerical difference $\delta_{ij}(x') - \delta_{ij}(x)$. Another point of this paper is that we need to *first aggregate the information* after taking the first differences, and then take the second difference in the *orders*. An important special case for the linear index model is the linear probability model, when F is uniform (on a suitable support). Then the screening difference becomes actual double-differencing of the outcomes. Another interesting special case is when the F is logistic. Then we could double-difference the outcomes of a tetrad after conditioning that the signs of the first differences are switching in the quadruple. This is very close to the result of Graham (2017).

same event is true, the new realization of the double-differenced observable vector rules out another half-space. At the second picture of Figure 1.5 we can see that this rules out an additional segment of possible β-s, leaving us with a smaller set of possible coefficient vectors (the red segment on the unit circle). Figure 4 repeats this thought experiment to demonstrate how we are getting closer and closer to β as we take more and more couples for which the G_{ij} event is true. As Shum et al (2017) also shows[7], this kind of identification argument is very similar to the semiparametric idenfitication proofs in Manski (1985, 1987, 1988). However, since $G_{ij} \Leftrightarrow \Delta_{ij}(x)\beta > A_j - A_i > \Delta_{ij}(x')\beta$, in addition we need to make sure that this event happens with positive probability, and that we have the appropriate sufficient variation conditions on the observables for identification (as for example in Han 1987), simultaneously.

Due to the similarities, in this section we will follow the language in Manski (1985) to prove an identification result.[8] For this we need some additional definitions. Given a b vector on the unit sphere such that $b \neq \beta$, in order to differentiate between b and β, on the one hand we need to have realized values of X_i, X_j such that

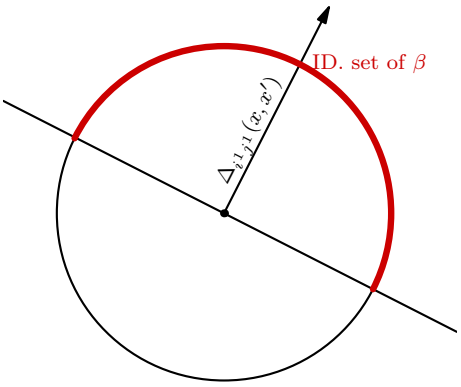
$$\text{sgn}(\Delta_{ij}(x) - \Delta_{ij}(x'))b \neq \text{sgn}(\Delta_{ij}(x) - \Delta_{ij}(x'))\beta. \tag{1.9}$$

On the other hand, it also should be true that

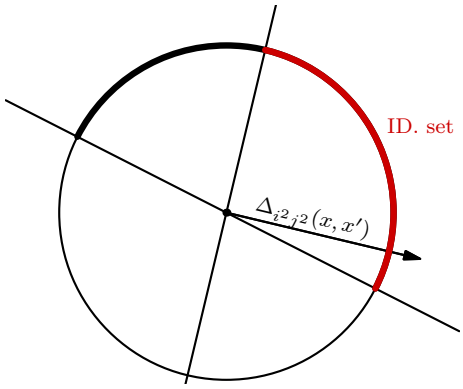
$$\Delta_{ij}(x)\beta > A_j - A_i > \Delta_{ij}(x')\beta \Leftrightarrow G_{ij} \tag{1.10}$$

[7]Independently, they have the same figures as Figure 1.5 in a slightly different setting (they only need to difference once).

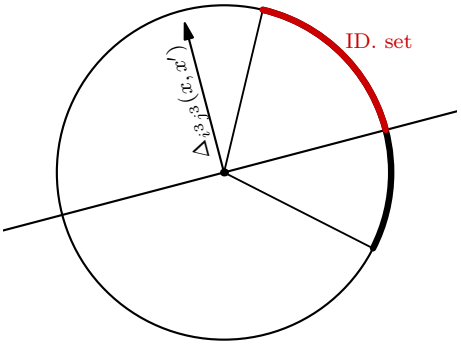
[8]A lemma saying that the population equivalent of our estimator identifies the coefficient vector up to scale is proven independently from this result, so if the Reader is only interested in the validity of the estimator, they can skip to the next subsection without any loss.



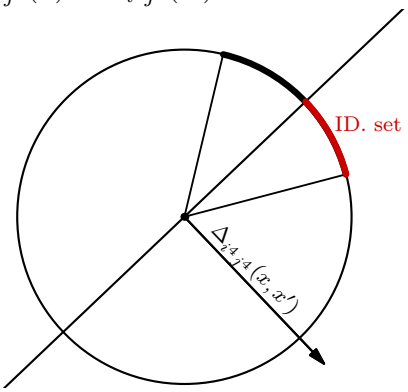
(a) Information after the first realization of double-differences $\Delta_1 = \Delta_{i^1 j^1}(x) - \Delta_{i^1 j^1}(x')$.



(b) Information after the second realization of double-differences $\Delta_2 = \Delta_{i^2 j^2}(x) - \Delta_{i^2 j^2}(x')$.



(c) Information after the third realization of double-differences $\Delta_3 = \Delta_{i^3 j^3}(x) - \Delta_{i^3 j^3}(x')$.



(d) Information after the fourth realization of double-differences $\Delta_4 = \Delta_{i^4 j^4}(x) - \Delta_{i^4 j^4}(x')$.

Figure 1.5: Information content of the double-differences of the observables if we observe that the screening values switch signs. Given that we observe that for the four pairs of nodes $(i^1, j^1), (i^2, j^2), (i^3, j^3), (i^4, j^4)$ the screening values switch signs, the data imply that the true β is on the red segments on the first, second, third and fourth unit circles as we cumulatively consider the first, second, third and fourth pair of nodes. The subfigures show how this red segment shrinks as we get more and more couples.

at the same time (jointly) with a positive probability. Therefore, for $x, x' \in Int\,(Supp(X_i))$, define

$$Z_{ij} = (\Delta_{ij}(x)\beta, \Delta_{ij}(x')\beta, A_i, A_j) \tag{1.11}$$

and

$$\begin{aligned} Z_b = \{(d_1, d_2, a_1, a_2) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R} \times \mathbb{R} : &(d_1 - d_2)b < 0 < (d_1 - d_2)\beta \text{ AND} \\ &\text{AND } d_1\beta > a_1 - a_2 > d_2\beta\}. \end{aligned} \tag{1.12}$$

We say that β is identified if for any b

$$P(Z_{ij} \in Z_b) > 0. \tag{1.13}$$

The following lemma states identification.

Lemma 2 (Identification of coefficients.)

Given Assumptions 1-6, β is identified.

Proof. It is enough to prove that

$$P[(\Delta_{ij}(x) - \Delta_{ij}(x'))b < 0 < (\Delta_{ij}(x) - \Delta_{ij}(x'))\beta] > 0 \ \forall b, \tag{1.14}$$

and that

$$\begin{aligned} &P[\Delta_{ij}(x)\beta > A_j - A_i > \Delta_{ij}(x')\beta | X_i = x_i, X_j = x_j] + \\ &+P[\Delta_{ij}(x)\beta < A_j - A_i < \Delta_{ij}(x')\beta | X_i = x_i, X_j = x_j] > 0 \ \forall x_i, x_j \in Supp(X_i) \end{aligned} \tag{1.15}$$

by the relationship between conditional and joint probabilities and symmetry of the inequalities above.[9]

For the first part, the set of admissible $\Delta_{ij}(x) - \Delta_{ij}(x')$ vectors are in a cone that is defined by one half of the intersect of two open half-spaces. As such, this cone does not lie in a proper subspace of $\mathbb{R}^p$ (because it is open and non-empty, hence it needs to contain an open ball in $\mathbb{R}^p$). However, since it is a cone that originates from zero, it is not Lebesgue-measure zero (it has a volume in $\mathbb{R}^p$) and it represents some directions that correspond to a (non-zero) surface of a spherical wedge in $\mathbb{R}^p$. From the arguments in the Appendix, it follows that under our standing assumptions the support of the double-differenced observable values contains some ϵ-ball around zero. This implies that the probability that $\Delta_{ij}(x) - \Delta_{ij}(x')$ is in some spherical wedge (arbitrary small but has a volume) is always non-zero.

To illustrate that the desirable cone has non-zero volume even for $p > 2$, Figure 1.6 shows for $p = 3$ the cone (red) that we want the double-differences of the observables to fall into if we would like to differentiate between the b and β vectors in the picture. The gray sphere is the unit sphere in $\mathbb{R}^3$, the blue plane is the points perpendicular to β, while the green space is the points orthogonal to b.

[9]The proof of the tetrad inequality identification lemma below offers more detail on these issues if the Reader is interested.

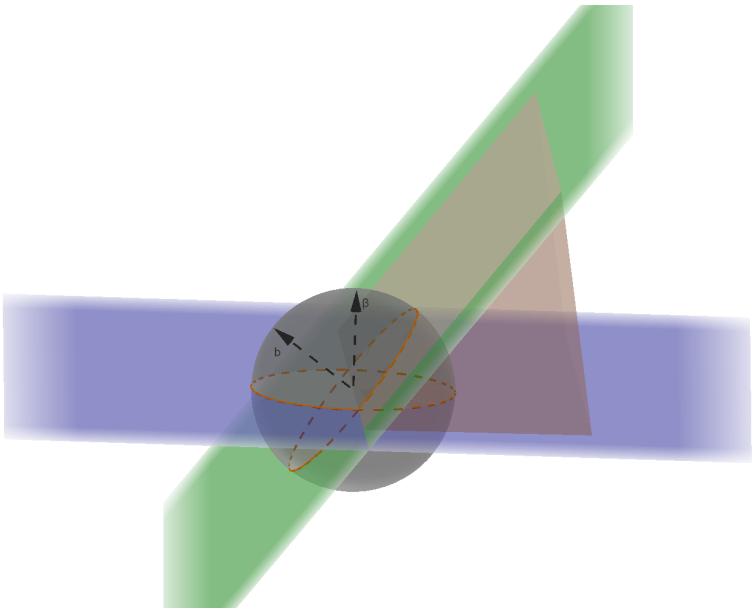


Figure 1.6: The figure demonstrates for $p = 3$ the volume (red cone) that we want the double-differences of the observables to fall into if we would like to differentiate between the b and β vectors in the picture. The gray sphere is the unit sphere in $\mathbb{R}^3$, the blue plane is the points perpendicular to β, while the green space is the points orthogonal to b.

In particular, we can get the first condition easily if we require

$$\begin{aligned} \Delta_{ij}(x)b < 0 < \Delta_{ij}(x')b \\ \Delta_{ij}(x)\beta > 0 > \Delta_{ij}(x')\beta. \end{aligned} \tag{1.16}$$

Given that there is a convex set on the support of W_{ij} with a non-empty interior, by the very same arguments as above (and using that x and x' is on the interior of the support), there is a positive probability that we find such x_i^*, x_j^* couple that satisfy the equations in 1.16. Conditioning on any of these couple, since by Assumption 5

the zero is on the conditional support of $A_j - A_i$, the probability

$$P[|A_j - A_i| \subset [\Delta_{ij}(x')\beta, \Delta_{ij}(x)\beta]|X_i = x_i^*, X_j = x_j^*] > 0.$$

After integrating up over x_i^*, x_j^*, we conclude that the first condition and the second condition (equations 1.14, 1.15) are satisfied with positive (non-zero) probability.

□

1.3.2.3 Tetrad inequality identification

In this section we introduce the hidden tetrads behind our identification argument, and prove the identification lemma that is used to show the consistency of the tetrad inequality estimator.

It is easy to see that another version of the Lemma 2 in the previous section could be that for $x, x' \in Supp(X_i)$ and $\{(X_l, A_l)\}_{l=i,j}$ i.i.d. random vectors the true β value solves

$$\underset{b\in\mathbb{R}^p:||b||=1}{\arg\max}\ E[(\text{sgn}(\delta_{ij}(x)) - \text{sgn}(\delta_{ij}(x)))\text{sgn}(\Delta_{ij}(x) - \Delta_{ij}(x'))b]. \tag{1.17}$$

Here we observe something peculiar. It has been emphasized in Graham (2017), that one should use tetrads to consider all the information we have in the parametric version of our model. This also parallels the intuition from two-way fixed effect linear panel data models. When comparing the node i to j in a linear probability model, we would have two fixed effect values to cancel out, which would prompt the researcher to do double-differencing, requiring a tetrad. However, we did not use a quadruple of nodes in our identification argument, seemingly. The

other two legs of the tetrads are hidden in the screening process and the x and x' observable values. In addition to this, we may use any x, x' values on the interior of the support of X_i for defining the screening sets - how should we choose? To gather all information, we want to use *every* admissible pair of grid points. If we apply the natural probabilistic weighting (the marginal F_X), we get expectations of the tetrad inequalities. Define for $(X_l, A_l)_{l\in\{i,j,k,k'\}}$ i.i.d. random vectors of characteristics

$$Q^{TI}(b) = E\left\{(\text{sgn}[\delta_{ij}(X_k)] - \text{sgn}[\delta_{ij}(X_{k'})]) \cdot \text{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))b]\right\}. \quad (1.18)$$

Also, let us denote the complement of the open ϵ-ball around β defined by

$$\bar{B}_\epsilon = \bar{B}_\epsilon(\beta) = \{b \in \mathbb{R}^p : ||b|| = 1, ||b - \beta|| \geq \epsilon\}.$$

For this new statistic we have the following identification result.

Lemma 3 (Tetrad inequality identification)

Given Assumptions 1-4, $Q^{TI}(b)$ identifies β. That is, for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\sup_{b\in\bar{B}_\epsilon} Q^{TI}(b) \leq Q^{TI}(\beta) - \delta$$

Proof. The proof is included in the Appendix, and it involves the following steps. *First* we prove that $Q^{TI}(\beta) - Q^{TI}(b) > 0$ a.s. for all $b \in \bar{B}_\epsilon$. This is done as follows:

1. we show that the integrand is maximized by β for tetrads that have $\delta_{ij}(X_k) \neq \delta_{ij}(X_{k'})$,

2. we prove that under the support assumptions, there are non-zero measure of tetrad realizations that satisfy the condition,

3. we point out that even after the conditioning, the $\Delta_{ij}(X_k)$ vector still spans every direction in $\mathbb{R}^p$.

Second, we argue that $Q^{TI}(b)$ is continuous in b and $\bar{B}_\epsilon$ is compact, so there must be a direction in $\bar{B}^\epsilon$ that maximizes $Q^{TI}(b)$. Therefore, there is a b^* for which after defining $\delta = Q^{TI}(\beta) - Q^{TI}(b^*)$, we have that $\min_{b \in \bar{B}_\epsilon}[Q^{TI}(\beta) - Q^{TI}(b)] = \delta > 0$ by the pointwise result from the first part. □

1.3.3 Identification of fixed effects in the linear index model

In this section we will denote the realization of the unobserved characteristics random variable A_i (fixed effects) for the node i in the network as a_i. In this section, we view these realizations as *parameters*. This means that implicitly we are looking at a smaller probability space,[10] when talking about the fixed effects as parameters. In our model, we can summarize the identification results as follows. Given the linear index assumption, for two individuals $i, j \in \{1, 2, ..., N\}$, we can identify the ratio of the differences in their fixed effect *parameters* and the $||\beta||$, the overall effect of the observable part of the index. That is, the fixed effects inherit the scale normalization from the identification argument for the slope coefficient, but we also need to introduce a location normalization for them. Our preferred example for this normalization is selecting an anchor node (say, the first one), and normalizing its fixed effect parameter to zero. Another possibility would be to normalize the median of the A_i to zero in the original problem. As $N \to \infty$, we can identify the

[10] Technically, this can be viewed as conditioning, and that we 'identify' $a_i - a_j$ after conditioning on $A_i = a_i, A_j = a_j$.

median node, and set its fixed effect parameter to zero. The identification argument in this section can be applied to many models that specify the trade-off between the fixed effect parameters and some varying observables in a limited dependent variable model.

The following result assumes that the β is already identified. To identify every nodes parameter, we need that the variation in the differences in fixed effects is smaller than the variation in the observable index (the $\Delta_{ij}(X_k)\beta$). This is a necessary assumption as we learn of the deviations in a_i from the deviation in the index (see screening lemma). The next assumption means that for identification we assume that the observables give the larger share of the variation in the connecting probabilities.

Assumption 7 (Homophily can be overwhelming)

For every (a_i, x_i) *and* (a_j, x_j) *both on* $Supp((A_i, X_i))$ *we have*

$$\alpha_{ij} = a_j - a_i \in [-|W_{ij}\beta|, |W_{ij}\beta|],$$

or that there exists $x_k \in Supp(X_i)$*, such that*

$$Supp(A_j - A_i) \subset [-|W_{ik}\beta|, |W_{ik}\beta|] \cap [-|W_{jk}\beta|, |W_{jk}\beta|]$$

The intuition of the identification lemma is summarized in Figure 1.7 for the case when $p = 2$. If the screening value for i and j with respect to a screening set x^* is exactly zero, then by the screening lemma the difference in the observable part of the indices exactly offset the differences in the fixed effect parameters. In Figure 1.7 we see that nodes i and j have the same probability of connecting with

a random node from the set of nodes with observable vector x^*. By the screening lemma this implies that $w(x_i, x^*)\beta - w(x_j, x^*)\beta = a_j - a_i$, where the left-hand side is identified.

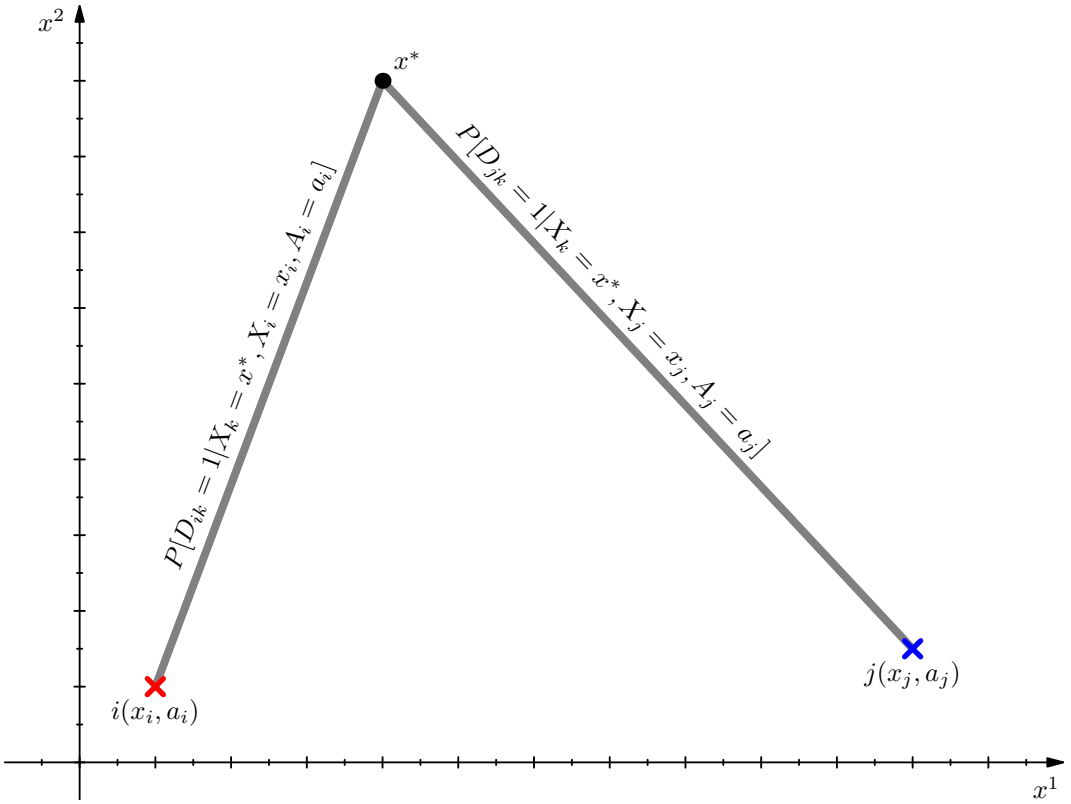


Figure 1.7: The figure shows the nodes i, j and the observable values x^* defining the screening set on the plane of observables. The width of the segments connecting the points is equal, signifying that the connection probabilities between i, j and a random node from the screening set (nodes with observable vector x^*) are the same. That means that the effect of the different homophily terms exactly offset the differences in the fixed effect parameters.

Lemma 4

Let Assumptions 1-7 hold. Then if $a_j - a_i \in [-|W_{ij}\beta|, |W_{ij}\beta|]$*, then*

$$a_j - a_i = E[\Delta_{ij}(X^*)\beta | \delta_{ij}(X^*) = 0, A_i = a_i, A_j = a_j].$$

If not, then there exist a node with some characteristics vector (x_k, a_k)*, for which* $a_i - a_k$ *and* $a_j - a_k$ *is identified as above, so that* $a_j - a_i$ *as well.*

Proof. Please find the proof in the Appendix. The proof based on the observation of the screening lemma that

$$0 = \delta_{ij}(x) \Leftrightarrow \Delta_{ij}(x)\beta = A_j - A_i. \tag{1.19}$$

In addition, $[-|W_{ij}\beta|, |W_{ij}\beta|] \subseteq Supp(\Delta_{ij}(X_k)\beta | X_i = x_i, X_j = x_j)$, since $X_k = x_i$ and $X_k = x_j$ is on the support, so then by assumption $\Delta_{ij}x_j = W_{ij}$ and $\Delta_{ij}x_i = -W_{ij}$ are both on the conditional support of $\Delta_{ij}(X_k)|X_i = x_i, X_j = x_j$. Moreover, because the support of the X_i is convex, and the distance function $w()$ is continuous, we have that the whole interval $[-|W_{ij}\beta|, |W_{ij}\beta|]$ is on the support of first differences by the intermediate value theorem. Note that if the distance function obeys the triangle inequality (so we have metrics for real), this bound is sharp. □

That is, if we fix the characteristics vector of i and j, and screen with an observational group x_k such that the screening value is 0, then $\Delta_{ij}(x_k)\beta$ gives the differences in the fixed effects of i and j. The first condition of Assumption 7 in this section assumes that such an x_k exists. Since there may be multiple such x_k values, we take expectations over them (using the marginal of X_i).

However, when for example $x_i = x_j$, then $Supp(\Delta_{ij}(X_k)\beta) = 0$ by assumption, so we can only identify the *ordering* of a_i, a_j from the argument corresponding

to Figure 1.7. The second part of Assumption 7 assumes that in this case there is a third node k that is 'far away' enough from both i and j (in terms of observables), such that $a_i - a_k$ and $a_j - a_k$ are both identified. Then since $a_j - a_k - (a_i - a_k) = a_j - a_i$, the difference α_{ij} is identified as well.

The resulting equality in the identification lemma above will be the basis of the estimator in the next section.

Remark. {If $a_j - a_i > \max_d[d \in Supp(W_{ij}\beta | X_i = x_i, X_j = x_j)] \geq 0$, then $\delta_{ij}(X_k) > 0|_{X_i=x_i, X_j=x_j}$ almost surely. And also, if $a_j - a_i < \min_d[d \in Supp(W_{ij}\beta | X_i = x_i, X_j = x_j)] \leq 0$, then $\delta_{ij}(X_k) < 0|_{X_i=x_i, X_j=x_j}$ almost surely. That is, we can always provide a lower/upper bound for $a_j - a_i$, and we can always identify the sign of the difference. }

As we mentioned above, the FE identification lemma only identifies the FE-differences. This means we can normalize the fixed effect parameter of a fixed node to zero, which is a sufficient normalization to identify the FE-s parameters for the rest of the network. To finish our identification analysis of the whole model, we need to remember that we only identify the β up to scale normalization $||\beta|| = 1$, so this implicitly inherited by the FE-identification problem. We get that the FE-parameters are identified up to scale and location normalizations.

1.4 Estimation

1.4.1 Tetrad inequality estimator

The tetrad inequality estimator is based on the identification result with $Q^{TI}(b)$.

Definition 1

$$\hat{\beta}^{TI} = \underset{b\in\mathbb{R}^p:||b||=1}{\arg\max}\ \hat{Q}_n^{TI}(b).$$

Here $\hat{Q}_n^{TI}$ *is defined as*

$$\hat{Q}_n^{TI} = \binom{n}{2}^{-1}\binom{n-2}{2}^{-1}\sum_{i<j,k<k'}\left[sgn[\hat{\delta}_{ij}(X_k)] - sgn[\hat{\delta}_{ij}(X_{k'})]\right]\cdot \\ \cdot\, sgn\left[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))b\right]$$

where $k, k' \neq i, j$

$$\hat{\delta}_{ij}(x) = \frac{\sum_l (D_{il} - D_{jl}) K\left[\frac{X_l - x}{\sigma_n}\right]}{\sum_l K\left[\frac{X_l - x}{\sigma_n}\right]}.$$

$K : \mathbb{R}^p \to [0,1]$ *is a kernel function and* σ_n *is a bandwidth sequence that is going to zero as* n *approaches infinity with a slower rate than* $n^{-1/p}$.

$\hat{Q}_n^{TI}(b)$ is defined as a statistic that is similar to a fourth-order U-statistic, except that the $\hat{\delta}_{ij}(.)$ are estimated. The presence of the $\hat{\delta}_{ij}(k)$ estimators makes the problem more complicated for two reasons. First, the kernel estimates contain the U_{ij}-s, that are drawn for every possible pair, not for the original unit of observation (the nodes). Second, there is another level of aggregation that involves the D_{ij}-s, and so indirectly the *whole* (X, A) sample is included in every term of $\hat{Q}_n^{TI}(b)$.[11]

1.4.1.1 Consistency of the tetrad inequality estimator

In this section we prove the (weak) consistency of the tetrad inequality estimator. We will follow Pakes and Pollard (1989) for the consistency proof. The

[11] Technically, this makes this statistic an infinite-order U-statistic.

identification lemmas have already established that if we have an arbitrarily good estimator for $Q^{TI}(b)$, then with the help of that statistic, we can rule out all candidates for the β vector that are at least $\epsilon > 0$ far away from the true value. So in principle, all we need to show is the uniform convergence[12] of $\hat{Q}_n^{TI}(b)$ to $Q^{TI}(b)$. For all the results here we assume random sampling.

Assumption 8 (Random sampling.)

The researcher either observes the whole network, or randomly samples the nodes into the observed network.

Known $\delta_{ij}(x_k)$-s

Assume that we observe the true screening values. If we know them, $\delta_{ij}(X_k)$ is just a function of $X_i, X_j, A_i, A_j, X_k, X_{k'}$. Define

$$\begin{aligned}\tilde{Q}_n^{TI}(b) = &\binom{n}{2}^{-1}\binom{n-2}{2}^{-1}\sum_{i<j,k<k'}(\text{sgn}[\delta_{ij}(X_k)] - \text{sgn}[\delta_{ij}(X_{k'})])\cdot \\ &\cdot \text{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))b] = \\ &= \binom{n}{4}^{-1}\sum_{i<j<k<k'} l(X_i, X_j, X_k, X_{k'}, A_i, A_j),\end{aligned} \tag{1.20}$$

where the function l is the symmetrized kernel of the forth-order U-statistic corresponding to $\tilde{Q}_n^{TI}(b)$. The symmetrization is always possible as described in Serfling (1980) Chapter 5. Since it is a sum of six terms of evaluations of the asymmetric kernel, it will inherit the boundedness property.

[12] over the b-s

Define

$$\tilde{\beta}^{TI} = \arg\max_{||b||=1} \tilde{Q}_n^{TI} \tag{1.21}$$

Lemma 5

Under Assumption 1-8, $\tilde{\beta}^{TI}$ consistently estimates β.

Proof. The proof in the Appendix follows standard U-statistic theory, and adopts a similar approach to Sherman (1994). □

The effect of $\delta_{ij}(X_k)$ being estimated

Because a kernel estimator is used for our first stage (screening), we borrow assumptions from the literature of the Nadaraya-Watson estimator (for example Hansen 2008).

Assumption 9 (Nadaraya-Watson estimator)

The following conditions hold on the kernel and the bandwidth in the definition of $\hat{\beta}^{TI}$:

- *The kernel K is a Parzen-Rosenblatt kernel that is twice continuously differentiable.*
- *The (X_i, A_i) vector is supported on a compact set. Moreover, the joint pdf exists, and it is bounded away from zero and infinity.*
- *The bandwidth σ_n goes to zero with $n \to \infty$, but $n\sigma_n^p \to \infty$.*

Remark. {The compact support assumption of the characteristic vector can be weakened to tail assumptions, as it is customary in some part of the literature. If we further restrict the kernel choice, we can weaken the smoothness assumption on the F as well.}

We need to prove that

Lemma 6

Under Assumptions 1-9,

$$\lim_{n\to\infty} \sup_{||b||=1} E[|\hat{Q}_n^{TI}(b) - \tilde{Q}_n^{TI}(b)|] = 0. \tag{1.22}$$

If this is true, then consistency of $\hat{\beta}^{TI}$ follows by the same argument we used to prove the previous lemma. The proof is included in the Appendix.

Proposition 1

Under Assumptions 1-9, plim $\left(\hat{\beta}^{TI}\right) = \beta$.

Proof. The proof of convergence in Lemma 5 and Lemma 6 proves uniform convergence for $\hat{Q}_n^{TI}(b)$ after using the triangle inequality

$$\sup_b |\hat{Q}_n^{TI}(b) - Q^{TI}(b)| \leq \sup_b |\tilde{Q}_n^{TI}(b) - Q^{TI}(b)| + \sup_b |\hat{Q}_n^{TI}(b) - \tilde{Q}_n^{TI}(b)|.$$

Further steps could include the again the standard arguments from Newey and McFadden (1994) using the identification lemma (Lemma 3) and uniform convergence. □

1.4.2 The simplified inequality estimator

In this section we introduce the simplified version of the tetrad inequality estimator, and examine some of its properties.

The simplified tetrad inequality estimator is defined as

Definition 2

$$\hat{\beta} = \arg\max_{||b||=1} Q_n(b),$$

where

$$Q_n(b) = \binom{n}{2}^{-1} \sum_{i<j} (sgn[\hat{\delta}_{ij}(x)] - sgn[\hat{\delta}_{ij}(x')]) \cdot sgn[(\Delta_{ij}(x) - \Delta_{ij}(x'))b], \quad (1.23)$$

where x and x' are fixed (non-random) vectors on the interior of the support of X_i.

In addition, we will sometimes use the shorthands $F_{ik} = F[W'_{ik}\beta + A_i + A_k]$ and $\Delta_{ij} = \Delta_{ij}(x) - \Delta_{ij}(x')$.

The $\hat{\beta}$ estimator is computationally and conceptually simpler than $\hat{\beta}^{TI}$. It does not require to add up the score for every possible tetrad from the network, and the objective function is very similar to a second-order U-process (as opposed to a 4th-order one). This variation of the tetrad inequality estimator may also be attractive from the point of view of the empirical researcher, as the researcher does not need to see the adjacency matrix to calculate the estimator. One would only need to randomly select nodes and ask them about how many friends they have with the previously chosen observables vector. This may decrease the burden of

data collection significantly. This estimator makes it possible to use more aggregate data from possibly different sources for estimation as well.

By the same identification argument as in Lemma 3 for the tetrad inequality estimator, under the same assumptions, the β is identified as the maximizer of

$$Q(b) = E[(\text{sgn}\delta_{ij}(x) - \text{sgn}\delta_{ij}(x'))\text{sgn}(\Delta_{ij}(x)b - \Delta_{ij}(x')b)]. \tag{1.24}$$

Also, by exact analogue arguments to the proofs of Lemma 5-6, $Q_n(b)$ converges uniformly to $Q(b)$, given Assumptions 1-9. This means that the simplified tetrad inequality estimator is consistent by the same argument as the tetrad inequality estimator.

1.4.2.1 Asymptotic theory of the simplified tetrad inequality estimator

In this section we argue that the simplified tetrad inequality estimator is converging with a non-parametric rate. The main steps to analyze this estimator are

1. calculating the mistake probability of screening,
2. verifying the Hoeffding-decomposition for our objective function and
3. using the argument from Sherman (1994) to determine the rate of convergence of the estimator.

Mistake probability of screening

Our first step is to see what is the (unconditional) probability that we miss the right sign of the screening value, that is, $P[\text{sgn}\hat{\delta}_{ij}(x) \neq \text{sgn}\delta_{ij}(x)]$. This probability turns out to converge to zero with a non-parametric rate.

Lemma 7 (Rate of mistake probability)

Given Assumptions 1-9 and that f (the pdf corresponding to F) is bounded away from zero and infinity, for $|\delta_{ij}| > 0$

$$\begin{aligned} &E[|sgn\hat{\delta}_{ij}(x) - sgn\delta_{ij}(x)||X_i, X_j, A_i, A_j] = \\ &= P[|sgn\hat{\delta}_{ij}(x) - sgn\delta_{ij}(x)| > 0|X_i, X_j, A_i, A_j] \\ &= O(\exp(-n\sigma_n^p)), \end{aligned}$$

however,

$$E[|sgn\hat{\delta}_{ij}(x) - sgn\delta_{ij}(x)|] = O(\sqrt{n\sigma_n^p}^{-1})$$

Proof. The proof is in the Appendix. Given the uniform convergence rates of the Nadaraya-Watson estimator, this may not be a surprising result. After the rewriting

$$\begin{aligned} &P[|\text{sgn}[\delta_{ij}(x)] - \text{sgn}[\hat{\delta}_{ij}(x)]| > 0|X_i, X_j, A_i, A_j] \leq \\ &\leq P[|\hat{\delta}_{ij}(x) - \delta_{ij}(x)| > |\delta_{ij}(x)||X_i, X_j, A_i, A_j], \end{aligned} \tag{1.25}$$

a small modification of the arguments in for example Andrews (1994b) and Audibert and Tsybakov (2007) gives the proof. □

Hoeffding-decomposition

In this section we will denote $Z_i = (X_i, A_i)$. We also set aside the way we get the screening values, and introduce $\tau(\omega) : (Z_i, Z_j) \to \{-1, 0, 1\}$ random function, a

classification rule. For example, when using the Nadaraya-Watson estimator, we had

$$\tau(Z_i, Z_j) = 0.5[\text{sgn}\hat{\delta}_{ij}(x) - \text{sgn}\hat{\delta}_{ij}(x')].$$

Let us denote the support of the τ function as $\mathcal{T}$. Also define $\tau_0 : (Z_i, Z_j) \to \{-1, 0, 1\}$ as the true classification:

$$\tau_0(Z_i, Z_j) = 0.5[\text{sgn}\delta_{ij}(x) - \text{sgn}\delta_{ij}(x')]$$

Implicitly, we assume that $\tau_0 \in \mathcal{T}$, and we omit the n subscript for τ, but we are thinking of a sequence of τ-s (because we get more and more observations to screen with). To emphasize which τ (sequence) we are using to calculate the objective functions Q_n, from now on we will include it as an argument of said function.

Remark. {There are several other classification methods that may result in the condition that $\mathcal{T}$ is Euclidean. Unfortunately, our chosen classification method based on the kernel estimator does not result in such a rule. However, if one could enforce this condition on the classifier, then the classical results regarding the Hoeffding-decomposition would hold as in Sherman (1994) or Arcones and Gine (1991). }

By the Hoeffding-decomposition following Serfling (1984), we can write

$$Q_n(b, \tau_0) = \binom{n}{2}^{-1} \sum_{i<j} g + f(Z_i) - g + f(Z_j) - g + u^{ij}(Z_i, Z_j) = \tag{1.26}$$

$$= g + 2[n(n-1)]^{-1} \sum_i \sum_{i \neq j} [f(Z_i) - g] + \binom{n}{2}^{-1} \sum_{i<j} u^{ij}(Z_i, Z_j)$$

Where

$$f(Z_i) = f(Z_i; b) = E[\tau_0(Z_i, Z_j)\text{sgn}(\Delta_{ij}b)|Z_i] \tag{1.27}$$

$$g = g(b) = E[\tau_0(Z_i, Z_j)\text{sgn}(\Delta_{ij}b)] \tag{1.28}$$

$$u^{ij}(Z_i, Z_j) = u(Z_i, Z_j; b) = \tau_0(Z_i, Z_j)\text{sgn}(\Delta_{ij}b) - f(Z_i) - f(Z_j) + g. \tag{1.29}$$

Here $\tau_0(Z_i, Z_j)$, the pointwise evaluation of τ_0. Taking advantage of the fact that the second stage resembles an MRC estimator, we could use very similar arguments to Sherman (1994) to prove that under appropriate smoothness conditions on the distribution functions $Q_n(b, \tau)$ is driven by the first and second components of the decomposition, which are constant and behave like an empirical process, respectively.

However, in our case we have to account for estimating τ_0, so we need to modify 1.27. The following lemma states the decomposition with an additional discrepancy term $Q_n(b, \tau_0) - Q_n(b, \tau)$. However, the results to go through, we need a technical condition on the distribution of the Δ_{ij} and the U_{ij} first.

Assumption 10

Given that $||\Delta_{ij}|| > 0$, *the pdf of* $\frac{\Delta_{ij}}{||\Delta_{ij}||}$ *is bounded away from infinity.*

Moreover, the pdf f *(corresponding to the cdf* F*) is bounded away from zero and infinity on the support of* $W_{ij}\beta + A_i + A_j$.

Remark.{This assumption is satisfied if $w^i(x, y) = (x - y)^2$ or $w^i(x, y) = |x - y|$ for all i from the set $\{1, 2, .., p\}$, and the pdf of the observables is bounded away from zero and infinity on their support. The random variable $\frac{\Delta_{ij}}{||\Delta_{ij}||}$ has a

compact support. In general, the only additional condition we need to see that the pdf of this variable is continuous. }

Lemma 8 (Hoeffding-decomposition)

Under Assumption 1-10, if b is in an $o_p(1)$ neighborhood of β,

$$Q_n(b,\tau) - Q_n(\beta,\tau) = g(b) - g(\beta) + O_p(\sqrt{n}^{-1})O_p(||b-\beta||) + O_p(n^{-1}) + d_n$$

where

$$d_n(b,\tau) = \binom{n}{2}^{-1} \sum_{i<j} (\tau_{ij} - \tau_0(Z_i, Z_j))[sgn(\Delta_{ij} b) - sgn(\Delta_{ij}\beta)] \leq \tag{1.30}$$

$$\leq O_p\left(\sup_{(Z_i,Z_j)} |\tau_{ij} - \tau_0(Z_i, Z_j)| \right) O_p(||b-\beta). \tag{1.31}$$

Proof. The proof is in the Appendix. □

Huber-Pollard-Sherman lemma

Now we are ready to apply the arguments in the 'General Method' section of Sherman (1994). According the results due to Sherman, to get $\sqrt{n}$-consistency for the simplified tetrad estimator, we need that the expected bias of $\hat{\tau}$ is $O(\sqrt{n}^{-1})$. This is not going to be true for the $\hat{\tau}$ procedure we have currently, as it is expressed in the lemma about the rate of mistake probability, as we have seen that the bias is the same order as $\sqrt{n\sigma_n^p}^{-1/2}$, where σ_n is the bandwidth and $p \geq 2$ is the number of observables. Therefore the bias of the first stage will be a bottleneck for the convergence of the estimator. The result is summarized in the following proposition.

Proposition 2

Given Assumptions 1-10, the simplified tetrad inequality estimator $\hat{\beta}$ is approaching the true value β with a non-parametric rate

$$||\beta - \hat{\beta}|| = O_p\left(n^{-\frac{1}{2}+\frac{sp}{2}}\right)$$

for $O(\sigma_n) = n^{-s}, 0 < s < 1/p$.

Proof. See Appendix. □

The tetrad inequality estimator

The main aim of this section was to show that the simplified estimator is converging at a non-parametric rate, as we determined, and the Monte Carlo simulations also buttress this result. One can argue that the tetrad-inequality estimator will inherit the same problems that we saw earlier in this section. The full estimator has much more information, but we can write up a similar decomposition as in (1.92). We will still end up with a bias term that has a non-parametric rate, although given that we change the first two legs of the tetrads symmetrically, we achieve some bias reduction.

1.4.3 Estimator for the fixed effect-differences

We can define an estimator for the differences in the fixed effects based on the identification argument for the FE-s. Given a consistent estimator $\hat{\beta}$ for the β in the model, we can define an estimator for $\alpha_{ij} = a_j - a_i$ (the fixed effect *parameters*):

Definition 3

Fix the observables for i and j. Given that a $\hat{\beta}$ consistently estimating the true β, let

$$\hat{\alpha}_{ij} = \frac{\sum_{k \neq i,j} L\left[\frac{\hat{\delta}_{ij}(X_k)}{\sigma_n^l}\right] \Delta_{ij}(X_k)\hat{\beta}}{\sum_{k \neq i,j} L\left[\frac{\hat{\delta}_{ij}(X_k)}{\sigma_n^l}\right]},$$

where $L : \mathbb{R} \to \mathbb{R}$ is a Lipschitz-continuous, continuously differentiable, symmetric kernel function and $\sigma_n^l > 0$ is a sequence that is going to zero as n approaches infinity, but it decreases slower than n^{-1}.

The estimator is a sample analogue of the expectation in Lemma 4. The conditioning on a zero-probability event is implemented by a symmetric kernel.

The following proposition is summarizing what we know about the asymptotic theory of the fixed effect estimator.

Proposition 3 (Fixed effect estimator)

Fix nodes i, j. Given a vanishing sequence $r_n \geq \sqrt{n}^{-1}$ eventually, such that $\sup_{i,j,k} |\hat{\delta}_{ij}(X_k) - \delta_{ij}(X_k)| = O_p(r_n)$, and $||\hat{\beta} - \beta|| = O_p(r_n)$, define $\hat{\alpha}_{ij}$ as above.

If Assumption 1-7 hold, and $0 \in Supp(\delta_{ij}(X_k))$,

$$|\hat{\alpha}_{ij} - (a_i - a_j)| < O_p\left(\max\left\{\sigma_n^l, r_n(\sigma_n^l)^{-2}\right\}\right).$$

If $0 \notin Supp(\delta_{ij}(X_k))$, still

$$plim\ sgn[\hat{\alpha}_{ij}] = sgn[\alpha_{ij}].$$

Proof. The proof is included in the Appendix. □

That is, the estimator is consistent, and converges with a non-parametric rate, that is at least 1/3 of the rate of the $\hat{\beta}$-s rate. The reason of the decrease in the rate of convergence is that the argument of the conditioning kernel is itself estimated non-parametrically. Moreover, even if the $a_j - a_i$ is not directly identified, we can always consistently estimate the sign of this difference.

Once we can consistently estimate the differences of the fixed effects, we can either normalize the median to be zero, or simply take the first node in the sample, and normalize $a_1 = 0$, so that $\hat{a}_i = \hat{\alpha}_{1i}$. From there the distribution of fixed effects (up to the scale and location normalizations mentioned) is identified conditional on any $X_i = x_i$ value. This estimation may be a starting point of further analysis that examines how the unobserved fixed effects depend on the observable levels, so the estimator above may be used to generate a new set of dependent variables. The rate of convergence may be better for the differences in conditional expectations (versus the individual effect-differences).

1.5 Some extensions and additional topics

Our earlier remarks already mentioned that following examples seen in semi-parametric models with limited dependent variables, we can weaken the conditions on the distribution of the observables to allow for some discrete characteristics. Another simple generalization would be considering directed network. In this section we give two more related topics that can be addressed with the same empirical strategy as the benchmark model. Finally, we give some considerations for the case when the true coefficients are zero across all homophily dimensions, and address

the identification of F.

1.5.1 Partial identification

A crucial circumstance to note in the heuristics of Proposition 2 is that the rate is governed by the bias term of the screening error. But that comes from the necessity of kernel smoothing, so if there is no need for it, the rate of convergence will be parametric. This corresponds to the case when we do not have point identification, because the observable characteristics are all discrete (with finite support). In this section we assume everything in the identification lemma (Lemma 2), except that the observables are continuously distributed.

In a partial identification environment the argument in the identification lemma can be used to determine the identified set $H(\beta)$. In particular, if we observe the $\delta_{ij}(x)$ and $\delta_{ij}(x')$ screening values (for $x, x' \in Int(Supp(X_i))$, denote the set of possible values for β after observing n nodes by H_n. Now we only are interested in the event when $\delta_{ij}(x) > 0 > \delta_{ij}(x')$, so let us define

$$G_{ij} = \mathbb{1}[\delta_{ij}(x) > 0 > \delta_{ij}(x')].$$

Then for $p = 2$, for example

$$|H_n| = \sup_{b,b' \in H_n} |b - b'| = \inf_{(i,j),(k,l):G_{ij}=G_{kl}=1} \left|\left| \frac{\Delta_{ij}}{||\Delta_{ij}||} + \frac{\Delta_{kl}}{||\Delta_{kl}||} \right|\right|, \tag{1.32}$$

as it can be seen in Figure 1.5. That is, after selecting the (i, j) couples for which the G_{ij} identified event is observed, we form couples of these pairs and calculate the length of the sum of the differences of the observable vectors for every couple.

The resulting smallest number is a sharp bound for the diameter of the identified set.[13]

Claim 1

H_n is the directions corresponding to the dual cone of the conic average of these eligible observable difference vectors (for which $G_{ij} = 1$).

$$H_n = (co_{G_{ij}=1}(\Delta_{ij}))^*.$$

There are deterministic algorithms to generate the dual of a cone, and the normalization of the generators will give the identified set.

In turn, H can be defined as the limit of the H_n-s as $n \to \infty$. Define the finite set

$$S = \{s \in Supp\left(\frac{\Delta_{ij}}{||\Delta_{ij}||}|\Delta_{ij} \neq 0\right) : s\beta > 0\}, \tag{1.33}$$

the part of the support of the normalized double-differences that falls into the positive half-space defined by β as a kernel.

Claim 2

Under standing assumptions, the identified set for the direction of β is

$$H = (co(S))*, \tag{1.34}$$

where the $*$ superscript denotes the dual operation once again.

[13]The candidate coefficient vectors are normalized to length one to begin with.

The only problem is to determine whether the G_{ij} event happened, that is, the membership of some realization of the Δ_{ij} in S. However, screening behaves much better if we only need to take averages over the same observable value. This is feasible if we observe nodes born into the network with the same sets of observable characteristics infinitely often as n goes to infinity. Under our assumptions, the uniform mistake probability is diminishing with the rate $\sqrt{n}$.

1.5.2 The $\beta = 0$ case

In Assumption 4 we implicitly assume that $\beta \neq 0$. This is considered a nuisance case by the overwhelming part of the semiparametric literature for limited dependent variable models in the last 40 years. Indeed, for all applications we could consider, there is little doubt that there is a homophily effect in at least one dimension (typically distance). However, to close the identification analysis for the slope coefficients, we may want to examine the case if the screening values give an indication of the the value of $\mathbb{1}[\beta = 0]$.

While there may be better statistics for a basis of a test whether $\beta = 0$, our main observation here is that the sign of the screening values cannot vary with the screening set if and only if $\beta = 0$.

Claim 3

Given the assumptions in the identification lemma,

$$\beta = 0 \Leftrightarrow sgn\delta_{ij}(x) = sgn\delta_{ij}(x'), \forall (X_i, A_i), (X_j, A_j) \text{ and } x, x' \in Supp(X_i). \quad (1.35)$$

Proof. The direction that if $\beta = 0$ then the screening values do not change signs

while changing x, x' is trivial. As part of the proof for the identification lemma, we saw that if $\beta \neq 0$, then we can always vary x and x' enough such that $\Delta_{ij}(x)\beta > 0\Delta_{ij}(x')\beta$. Conditional on $X_i = x_i$ and $X_j = x_j$, we only need to make sure that x is closer to x_j, while x' is closer to x_i in every dimension (or vice versa, depending on the sign of the β-s). The measure of such x, x' couples is certainly non-zero under the marginal corresponding to X_i. Moreover, by the common support assumption regarding the FE-s (Assumption 5), we know that for any $\epsilon > 0$, $P[|A_i - A_j| < \epsilon | X_i = x_i, X_j = x_j] > 0$. This means that after conditioning on $X_i = x_i$ and $X_j = x_j$, the event

$$\Delta_{ij}(X)\beta > A_j - A_i > \Delta_{ij}(X')$$

has strictly positive (conditional probability).[14] Again, this is true for any x_i, x_j couple on the support, so by the monotonicity property of the integral we have that $P[\Delta_{ij}(X)\beta > A_j - A_i > \Delta_{ij}(X')]$. Now by the screening lemma we know that

$$\delta_{ij}(X) > 0 \Leftrightarrow \Delta_{ij}(X)\beta > A_j - A_i,$$

so under our assumptions, if $\beta \neq 0$

$$P[\text{sgn}\delta_{ij}(X) \neq \text{sgn}\delta_{ij}(X')] > 0.$$

□

This means that under our assumptions the observables vary enough (and there is enough mass of the fixed effects that does not) such that the screening

[14] Here X and X' are just iid draws from the distribution of X_i, just as in the data.

values switch signs with strictly positive probability when varying the screening sets, so the screening values can tell us if the true $\beta = 0$ or not. A similar procedure makes it testable if any element of the β vector is zero.

1.5.3 Identification of F

Claim 4

Given Assumptions 1-7, and the location normalization $Med(A_i) = 0$, the cdf of the unobservables, F is identified on the support of $W_{ij}\beta + A_i + A_j$.

Note that the support of the X_i-s is not restricted to a bounded set by these assumptions.

The argument for this claim does not suggest a good estimation procedure for F, but it is intuitive. Since we can identify the realized values of the FE-s and the β coefficients, we can identify the realized value of the index for any i, j in the population. This way we can plot out the

$$x = W_{ij}\beta + a_i + a_j \rightarrow E[D_{ij}] = F[x], \forall x \in Supp(W_{ij}\beta + A_i + A_j)$$

mapping. If the support of the index is assumed to be unbounded, we identify the whole function, if it is bounded, we only identify it on $Supp(W_{ij}\beta + A_i + A_j)$.

1.5.4 Information about the distance function

In both identification arguments, the fact that $0 \in Supp(\delta_{ij}(X_k))$ for a sufficiently large set of (X_i, X_j, X_k) played a pivotal role. In fact, as we argue in

the forthcoming chapter 2, one could regard the set

$$X_{ij}^* = \{x \in \mathbb{R} : \delta_{ij}(x) = 0\} \tag{1.36}$$

as the central object of the identification of this model. After fixing i, j nodes, by the screening lemma, for $x, x' \in Supp(X_i)$ we have

$$\Delta_{ij}(x) = \Delta_{ij}(x') = a_j - a_i \Leftrightarrow \delta_{ij}(x) = \delta_{ij}(x') = 0, \tag{1.37}$$

which means that the X_{ij}^* are part of the level curve of $\Delta_{ij}(x)$, as a function of x (after fixing X_i, X_j). Figure 1.8 depicts the definition of this set.

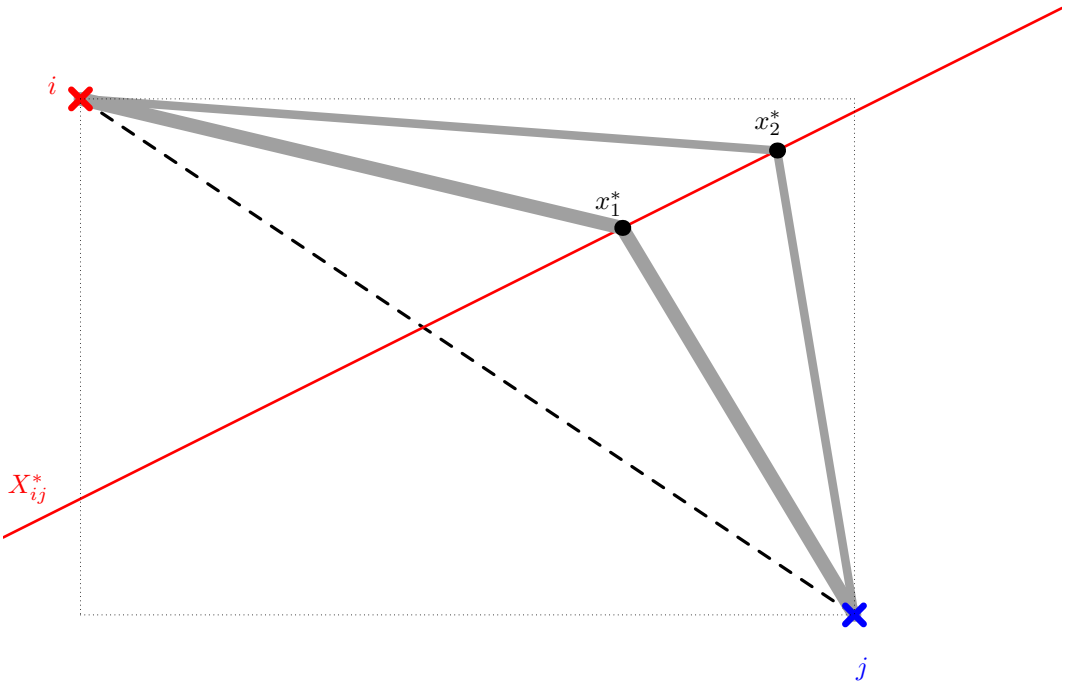


Figure 1.8: Definition of $X_{ij}^* = \{x \in \mathbb{R}^p : S_{ij}(x) = 0\}$, where $w(x_i, x)\beta - w(x_j, x)\beta = constant_{ij}$ is (part of) a level curve of $\Delta_{ij}(x)\beta$.

Then by the implicit function theorem, the slope of this level curve in some

direction is

$$-\frac{\frac{\partial w_1(x_j,x)-w_1(x_i,x)}{\partial x^1}}{\frac{\partial w_q(x_i,x)-w_q(x_j,x)}{\partial x^q}}, \tag{1.38}$$

which is exactly $\frac{\beta^1}{\beta^q}$, if the w_q and w_1 functions are the absolute value function, while it is $\frac{\beta^1(x_i^1-x_j^1)}{\beta^q(x_i^q-x_j^q)}$, if the $w_1(x,y) = w_q(x,y) = (x-y)^2$. This argument can be generalized to arbitrary polynomials.

Claim 5 (Distance function specification.)

Assume that for every $k = 1, ..., p$, $w_k(x,y) = (x-y)^{2c_k}$ *for* $c_k \in \mathbb{Z}^+$, *or* $w_k x, y = |x-y|$. *Then given Assumptions 1-6, then the specification of the distance function is refutable against the alternatives in this set.*

Based on this observation, chapter 2 answers the question if we can identify the distance function (test against any alternatives in a broad set).

1.6 Monte Carlo simulation

In this section we conduct a Monte Carlo simulation to illustrate the properties of the (simplified) tetrad inequality estimator. The distance function in our specification is the absolute value function, and $p = 2$. Screening should work relatively well, since the average probability of connecting is around 0.65, and the exogenous noise follows at first a logistic, then a Cauchy distribution. On the other hand, the fixed effect values highly depend on the observables in a linear (first dimension) and in a non-linear way (second dimension) as well, as even the support is changing somewhat with different X_i-s. The X_i-s are otherwise slightly

correlated and follow truncated normal distribution. The rate of the bandwidth is 0.33.

$$X_i^{1,2} \sim U[-2,2] \text{or} TN[-2,2,-1/0,2] \tag{1.39}$$

$$A_i = 0.25X_i^1 - 0.5|X_i^2| \cdot Z_i,\ Z_i \sim N_{[0,1]}[1,2] \tag{1.40}$$

$$U_{ij} \sim \text{logistic or Cauchy, or GEV-I} \tag{1.41}$$

$$D_{ij} = \mathbb{1}[\beta|X_i - X_j| + A_i + A_j \geq \epsilon_{ij}] \tag{1.42}$$

The choice of x and x' was *ad hoc*, after observing the support of the X_i-s. They are chosen to be in the interior, and far away from each other, but as it turns out they are not the best choice.[15] In the simulations below, the true value was 0.6.

First we run Graham's tetrad logit ($\hat{\beta}_{TL}$) and the tetrad inequality ($\hat{\beta}_{TI}$) for $n = 150$, which is a relatively low sample size. Graham fixes the distribution of the disturbance terms to be standard logistic, so we get a 2-vector of coefficients. In Table 1.1 we can see that even for the first dimension, for which the small sample bias is negligible, the misspecification bias is sizable when the true distribution of the unobservable noise is Cauchy. Interestingly, the small sample bias of the second dimension is higher, probably due to the nonlinearities introduced in the dependence between the FEs and this dimension of observables. However, the misspecification bias seems to be around 15% in each case. Since the distribution of the observables and the Cauchy distribution are both (close to) symmetric, we should not get much

[15]It seems that the bias can be substantially reduced if one chooses these points in a way that the distribution of the X_i is 'symmetric' around the points. However, these considerations are not the topic of present paper.

$\beta = (1, 0.6)$	logistic U_{ij}	Cauchy U_{ij}	GEV-I U_{ij}
$E[\hat{\beta}_{TL}]$	$(0.975, 0.526)$	$(0.854, 0.479)$	$(0.965, 0.435)$
$\sigma[\hat{\beta}_{TL}]$	$(0.035, 0.031)$	$(0.036, 0.034)$	$(0.040, 0.035)$
ratio of $\hat{\beta}_{TL}$	0.540	0.561	0.451
σ of ratio	0.034	0.044	0.036
$E[\hat{\beta}_{TI}]$	0.629	0.658	0.513
$\sigma[\hat{\beta_{TI}}]$	0.077	0.092	0.083

Table 1.1: Simulation results for Graham's tetrad logit and the tetrad inequality estimator ($\hat{\beta}_{TL}$) for $n = 150$. In the first column the true distribution follows the logistic distribution (no misspecification), in the second column the true distribution is Cauchy (misspecification), and the in third the GEV-I distribution. The number of repetitions is 500.

misspecification bias for the ratio of the coefficients. This property served as a sanity check during the simulations. However, when the misspecification included an asymmetric distribution for the noises, we could induce similar magnitude of misspecification bias even in the ratio. Our misspecification of choice was the generalized extreme value type I distribution (as defined by python/SciPy).

Comparing the statistics of the ratios is taking a conservative stance towards the performance of the tetrad inequality estimator, as we could refute the assumption behind the tetrad logit estimator that the variance of the disturbances is 1. On the one hand, comparing the biases of the individual estimates of Graham with the bias of the tetrad inequality estimator gives an overly optimistic view of our estimator. On the other hand, the multiplicative bias resulting from the Cauchy misspecification can be important from the point of view of the applied researcher (especially because the multiplying constant depends on the distribution of the observables). Even if we look at the coefficient ratios, the bias of the tetrad inequality

estimator is much lower for the asymmetric misspecification, and comparable in the two other cases. The standard deviation of the semiparametric estimator is roughly twice as high as that of the parametric estimator, which is to be expected. All in all, the semiparametric estimator performs better in terms of mean squared error in the third specification, even at this sample size.

The point of semi- and nonparametric results are that as we increase the sample size, the higher variance of the semiparametric estimator will decrease, but the (potentially) high bias of the parametric methods will remain. While this is certainly true for our case as well, in our network formation application we have two problems with this. First, increasing sample size does not necessarily result in proportionally more information in the data, second, the computational burden associated with calculating 4th-order U-statistics is going to be a problem. Here we only address the computational problem, with the introduction of the simplified estimator.

In particular, we run the *simplified* estimator for various sample sizes (Table 1.2). Understandably, these results are going to look worse, despite of the larger sample. We can see that the first stage disrupts the convergence rate of the estimator, as the bias of the simplified tetrad inequality estimator only decreases with a non-parametric rate. As we suspected from the theory, the variance behaves better, as the standard deviation decreases close to a $\sqrt{n}$ rate. This implies that the bias becomes relatively more and more important as we are increasing the sample size. Even if we calculate with the conservative 15% benchmark, the small sample bias of the simplified estimator will meet with the misspecification bias of the tetrad

logit estimator around the 500 sample size. If one assumes that the variance of the tetrad logit estimator is zero, the point where the simplified estimator is doing better than the tetrad logit will be somewhere between the $1,000$ and $2,000$ sample sizes (using the mean squared error as benchmark).

$\beta = 0.6$		logistic U_{ij} $\hat{\beta}$	Cauchy U_{ij} $\hat{\beta}$
N=250	mean	0.680	0.723
	sd. deviation	0.141	0.160
N=500	mean	0.670	0.694
	sd. deviation	0.084	0.097
N=1000	mean	0.647	0.684
	sd. deviation	0.053	0.061
N=2000	mean	0.637	0.672
	sd. deviation	0.034	0.040

Table 1.2: Monte Carlo simulation results for the simplified estimator with 1,000 replications for $p = 2$ (the first coefficient is normalized to one). The bandwidth is decreasing with 0.33 rate. The bias is decreasing with a nonparametric rate, the standard deviation is decreasing with a parametric rate.

In practice, to avoid the computational burden of the tetrad inequality estimator, the practitioner may consider two strategies. First, we could define the objective function of the simplified estimator for a finite set of (x, x') points. Even the introduction of an additional 3-4 of these points can achieve a better compromise between computation and efficiency. The second strategy is that for our estimator, the nodes we use in the first stage for screening do not have to be all included in the regression phase. The researcher could calculate the screening values only for a 1000 pair of i, j-s but could use all the observations available in the adjacency matrix for the first stage.

1.7 Conclusion

In this paper we suggested a new type of identification strategy for the average partial effects in a semi-nonparametric model of network formation with linear index. We defined an estimator based on the identification results, and proved its consistency. We also show that in this framework the distribution and in-sample levels of fixed effects are identified if normalized up to scale and location. For practical purposes, we suggested a simplified estimator that requires less computational time (in exchange for less information). We proved that the simplified estimator is converging at a non-parametric rate, and argued that the tetrad inequality estimator will inherit this property.

In forthcoming papers we present another identification strategy that can handle non-linear models as well, so that we can assume richer homophily effects in the same semi-nonparametric setting with degree heterogeneity. Moreover, we will show that it is possible to use the identification and estimation strategy of this paper to estimate the parameters of semi-parametric linear panel models with two-way fixed effects.

1.8 References

Albert, R., Barabasi, A. L. (2002). Statistical mechanics of complex networks. Reviews of modern physics, 74(1), 47.

Andrews, D. W.K. (1989). Asymptotics for Semiparametric Econometric Models: I. Estimation, Cowles Foundation Discussion Papers 908R, Cowles Foundation for Research in Economics, Yale University, revised Aug 1990.

Andrews, D. W. K. (1994a). Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity, Econometrica, Econometric Society, vol. 62(1), pages 43-72, January.

Andrews, D.W.K., (1994b). Nonparametric Kernel Estimation for Semiparametric Models, Econometric Theory, Cambridge University Press, vol. 11(03), pages 560-586, 1995 June.

Arcones, M. A., Gine, E. (1993,1991). Limit Theorems for U-Processes. Ann. Probab. 21 (1993), no. 3, 1494-1542.

Audibert, JY., Tsybakov, B. (2007). Fast learning rates for plug-in classifiers. Ann. Statist. 35 (2007), no. 2, 608-633.

Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding exponential random graph models. Annals of Statistics, 41(5): 2428–2461.

de Paula, Á. (2016). Econometrics of network models. Technical Report CWP06/16, CEMMAP.

de Paula, Á., Richards-Shubik, S., and Tamer, E. (2015). Identification of preferences in network formation games. Technical Report CWP29/15, CEMMAP.

Dzemski, A. (2014). An empirical model of dyadic link formation in a network with unobserved heterogeneity". Technical report, University of Gothenburg.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. Journal of Business and Economic Statistics, 31(3):253 – 264.

Graham, B. S. (2015). An econometric model of link formation with degree heterogeneity. Technical Report 20341, National Bureau of Economic Research. Published in Econometrica (2017).

Graham, B. S. (2015). Methods of identification in social networks. Annual Review of Economics, 7:465 – 485.

Graham, B. S. (2016). Homophily and transitivity in dynamic network formation (No. w22186). National Bureau of Economic Research.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. Journal of Econometrics, 35(2-3), 303-316.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. Econometric Theory, 24(03), 726-748.

Heckman, J. J. (2008). Econometric causality. International statistical review, 76(1), 1-27.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics, 19(3):293 – 325.

Jackson, M. O. (2008). Social and Economic Networks. Princeton University Press.

Jackson, M. O. and Wolinsky, A. (1996). A strategic model of social and economic networks. Journal of Economic Theory, 71(1):44 – 74.

Jochmans, K. (2017): Semiparametric Analysis of Network Formation,Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2017.1286242

Jochmans, K. and Weidner M. (2017): Fixed-Effect Regressions On Network Data, Working Paper

Kline, B. (2015). Identification of complete information games. Journal of Econometrics, 189(1), 117-131. Chicago

Lovasz, L. (2012). *Large Networks and Graph Limits.* AMS, Colloquium Publications, Vol 60.

Liu, N., Xu, H. (2012). Semiparametric analysis of social interactions with homophily. UT Austin Working Paper.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. Journal of Econometrics, 3(3):205 – 228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. Journal of Econometrics, 27(3):313 – 333.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. Econometrica, 55(2):357 – 362.

Mele, A. (2015). A structural model of segregation in social networks. Technical report, John Hopkins University.

Menzel, K. (2015). Strategic network formation with many agents. Technical report, New York University.

Nolan, D., and Pollard, D., (1987). U-Processes: Rates of Convergence. The Annals of Statistics, Vol. 15, No. 2 (June, 1987), 780-799.

Pakes, A., and Pollard, D. (1989). Simulation and the Asymptotics of Optimization Estimators. Econometrica, 57(5), 1027-1057.

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. John Wiley and Sons.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. Econometrica 61 123–137.

Sherman, R. P. (1994). Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators. Ann. Statist. 22

Sheng, S. (2014). A structural econometric analysis of network formation games. Technical report, UCLA.

Shi, X., and Shum, M. (2017). Estimating Semi-parametric Panel Multinomial Choice Models using Cyclic Monotonicity, WP

Snijders, T. A. B. (2011). Statistical models for social networks. Annual Review of Sociology, 37:131 – 153.

Stone, Ch. J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. Ann. Statist. 10 (1982), no. 4, 1040-1053.

1.9 Appendix

1.9.1 Preliminary claims

Claim 6

Take Assumption 2 and 6. Conditional on $X_i = x_i$ and $X_j = x_j$, where $x_i, x_j \in Int(Supp(X_i))$, there exists an ϵ-ball around zero $B_\epsilon(0) \in \mathbb{R}^p$ such that

$$B_\epsilon(0) \in Supp(\Delta_{ij}(X_k)|X_i = x_i, X_j = x_j).$$

Proof. We will use the property of the $w_k(.,.)$ functions that they are symmetric and continuous, and that for any $k \in \{1, .., p\}$ and $x, y \in \mathbb{R}$

$$x \neq y \Leftrightarrow w_k(x, y) \neq 0.$$

Fix an $X_i = x_i$ and an $X_j = x_j$. Then by Assumption 6, the vector

$$x_\lambda = \begin{pmatrix} \lambda^1 x_i^1 & +(1-\lambda^1)x_j^1 \\ \lambda^2 x_i^2 & +(1-\lambda^2)x_j^2 \\ & \cdots \\ \lambda^k x_i^k & +(1-\lambda^k)x_j^k \\ & \cdots \\ \lambda^p x_i^p & +(1-\lambda^p)x_j^p \end{pmatrix}$$

for $\lambda \in [0,1]^p$ is on the support of the X_i.

Define the functions $f_i : [0,1]^p \to \mathbb{R}^p$

$$f_i(\lambda) = w(x_i, x_\lambda),$$

and $f_j : [0,1]^p \to \mathbb{R}^p$

$$f_j(\lambda) = w(x_j, x_\lambda).$$

Both $f_i(\lambda)$ and $f_j(\lambda)$ are continuous functions, and so $f_i(\lambda) - f_j(\lambda)$ is continuous as well. Note in addition that

$$f_i^k(1) - f_j^k(1) = w^k(x_i^k, x_j^k) \tag{1.43}$$

$$f_i^k(0) - f_j^k(0) = -w^k(x_i^k, x_j^k). \tag{1.44}$$

Note that $p < \infty$, and the distances are calculated separably (independently across dimensions in W_{ij}). Then by the intermediate value theorem, if $0 < \epsilon < \min_k[w^k(x_i^k, x_j^k)]$, for[16] any $c \in [-\epsilon, \epsilon]^p$ there is a $\lambda_c \in [0,1]^p$ such that

$$f_i(\lambda_c) - f_j(\lambda_c) = w(x_i, x_{\lambda_c}) - w(x_j, x_{\lambda_c}) = c.$$

To close the argument we need that the ϵ-ball (the open $B_\epsilon(0)$) is on $[-\epsilon, \epsilon]^p$, and the w function is continuous. So after taking any point c from the ϵ-ball (open), we only need to look for its λ_c, and take a small enough δ-ball around x_{λ^c}, that will map back to the ϵ-ball:

$$\exists \delta > 0 : \cup_{\xi \in B_\delta(x_{\lambda_c})} w(x_i, \xi) - w(x_i, \xi) \subset B_\epsilon(0).$$

At the same time, $P[X_i \in B_\delta(x_{\lambda_c})] > 0$, since any x_λ was on the support of X_i. □

1.9.2 Screening lemma

In the linear index model with Assumption 1-4, for any $x, x' \in Supp(X_i)$

$$\delta_{ij}(x) > 0 \text{ and } \delta_{ij}(x') < 0 \Leftrightarrow \Delta_{ij}(x)\beta > A_j - A_i > \Delta_{ij}(x')\beta.$$

[16] Such ϵ exists by the metric properties of the $w_k(.,.)$-s. Also, technically, we should use the intermediate value theorem dimension-by-dimension, and then 'assemble' a λ_c vector, but since the w_k functions only have arguments from the kth elements of the observable vectors, the procedure is quite self-evident and omitted.

Proof. By definition, for $(X_i, A_i), (X_j, A_j)$, and $x \in Supp(X_i)$

$$\begin{aligned}\delta_{ij}(x) > 0 \Leftrightarrow E[D_{ik} - D_{jk}|X_k = x, A_i = a_i, A_j = a_j, X_i, X_j] > 0 \\ \Leftrightarrow E[\mathbb{1}[W'_{ik}\beta + A_i + A_k \geq u_{ik}]- \\ - \mathbb{1}[W'_{jk}\beta + A_j + A_k \geq u_{jk}]X_k = x, A_i = a_i, A_j = a_j, X_i, X_j] > 0.\end{aligned} \tag{1.45}$$

Because the pair-specific u_{ij}-s are exogenous and identically distributed, by the law of iterated expectation and the linearity of the expectation operator we have

$$\begin{aligned}\delta_{ij}(x) > 0 \Leftrightarrow ... \Leftrightarrow \\ \Leftrightarrow E[F[W'_{ik}\beta + A_i + A_k] - F[W'_{jk}\beta + A_j + A_k]|X_k = x, A_{i,j}, X_{i,j}] > 0.\end{aligned} \tag{1.46}$$

Due to the conditioning (we are screening two specific nodes all along with an observational group) and the i.i.d. assumption on the (X_i, A_i) vectors, the W_{ik}, W_{jk}, A_i and A_j are constants in the integrand. So regardless the realization $a = A_k$ takes, after conditioning by the strict monotonicity of F

$$F[W'_{ik}\beta + A_i + a] > F[W'_{jk}\beta + A_j + a] \Leftrightarrow W'_{ik}\beta + A_i > W'_{jk}\beta + A_j. \tag{1.47}$$

This means that the integrand in the previous conditional expectation is either exactly always positive or negative or zero, and this only depends on the sign of $W'_{ik}\beta + A_i - W'_{jk}\beta - A_j$. If we integrate (1.47) over a (using the measure $F_{A|x}$), we get

$$\begin{aligned}E[F[W'_{ik}\beta + A_i + A_k] > F[W'_{jk}\beta + A_j + A_k]|X_k = x, A_{i,j}, X_{i,j}] \Leftrightarrow \\ \Leftrightarrow W'_{ik}\beta + A_i > W'_{jk}\beta + A_j,\end{aligned} \tag{1.48}$$

by the monotonicity property of the integral. Note that this is only beautiful because of the linear index, since the a cancels out, and so we get

$$\delta_{ij}(x) > 0 \Leftrightarrow \Delta_{ij}(x)\beta > A_j - A_i. \tag{1.49}$$

On the LHS we have an identified event, on the RHS we got a relation between a function of β with the observables on the one hand, and the function of unobservables that is invariant to x_k on the other hand.

Note that the argument works in both directions, symmetrically for the case when $\delta_{ij}(x') < 0$. From (1.49) it follows the result:

$$\delta_{ij}(x) > 0 > \delta_{ij}(x') \Leftrightarrow \Delta_{ij}(x)\beta > A_j - A_i > \Delta_{ij}(x')\beta \tag{1.50}$$

□

1.9.3 Tetrad inequality identification

Given Assumptions 1-4, $Q^{TI}(b)$ identifies β. That is, for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\sup_{b \in \bar{B}_\epsilon} Q^{TI}(b) \leq Q^{TI}(\beta) - \delta$$

Proof. First we prove that $Q^{TI}(\beta) - Q^{TI}(b) > 0$ a.s. for all $b \in \bar{B}_\epsilon$. This is done as follows:

1. we show that the integrand is maximizied by β for tetrads that have $\delta_{ij}(X_k) \neq \delta_{ij}(X_{k'})$,

2. we prove that under the support assumptions there are non-zero measure of tetrad realizations that satisfy the condition,

3. we point out that even after the conditioning, by Claim 6, the $\Delta_{ij}(X_k)$ vector still spans every direction in $\mathbb{R}^p$.

Second, we argue that $Q^{TI}(b)$ is continuous in b and $\bar{B}_\epsilon$ is compact, so there must be a direction in $\bar{B}_\epsilon$ that maximizes $Q^{TI}(b)$. Therefore, there is a b^* for which after defining $\delta = Q^{TI}(\beta) - Q^{TI}(b^*)$, we have that $\min_{b \in \bar{B}_\epsilon}[Q^{TI}(\beta_0) - Q^{TI}(b)] = \delta > 0$ by the pointwise result from the first part.

PART 1: Pointwise result

Examining the integrand of $Q^{TI}(b)$

First, by the assumption that the observables are continuous random variables, the screening value will take the value zero with probability zero, so I ignore those cases. Therefore the first term in the integrand *almost surely* will take the value 2, whenever

$$\delta_{ij}(X_k) > 0 > \delta_{ij}(X_{k'}),$$

or -2, whenever

$$\delta_{ij}(X_{k'}) > 0 > \delta_{ij}(X_k),$$

or 0, when

$$\delta_{ij}(X_k), \delta_{ij}(X_{k'}) > 0 \text{ OR } \delta_{ij}(X_k), \delta_{ij}(X_{k'}) < 0.$$

By Lemma 1 then since the $\Delta_{ij}(X_k)\beta$-s are also continuously distributed,

$$\begin{aligned} \text{sgn}[\delta_{ij}(X_k)] - \text{sgn}[\delta_{ij}(X_{k'})] = 2 \Rightarrow \Delta_{ij}(X_k)\beta > \Delta_{ij}(X_{k'})\beta \ a.s., \\ \text{sgn}[\delta_{ij}(X_k)] - \text{sgn}[\delta_{ij}(X_{k'})] = -2 \Rightarrow \Delta_{ij}(X_k)\beta < \Delta_{ij}(X_{k'})\beta \ a.s., \end{aligned} \tag{1.51}$$

or to put it in another way, *given that* $\text{sgn}[\delta_{ij}(X_k)] \neq \text{sgn}[\delta_{ij}(X_{k'})]$,

$$\text{sgn}[\delta_{ij}(X_k)] - \text{sgn}[\delta_{ij}(X_{k'})] = 2 \ \text{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))\beta] \ a.s.,$$

As mentioned above, if $\text{sgn}[\delta_{ij}(X_k)] = \text{sgn}[\delta_{ij}(X_{k'})]$, then the integrand in $Q^{TI}(b)$ evaluates to zero.

From here, after using the law of total expectation it follows that

$$\begin{aligned} Q^{TI}(b) = 2 \ P[\text{sgn}[\delta_{ij}(X_k)] \neq \text{sgn}[\delta_{ij}(X_{k'})]] \cdot \\ \cdot E\left\{\text{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))\beta] \cdot \right. \\ \left. \cdot \text{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))b] | \text{sgn}[\delta_{ij}(X_k)] \neq \text{sgn}[\delta_{ij}(X_{k'})]\right\}. \end{aligned} \tag{1.52}$$

Again, because if the $\Delta_{ij}(X_k)$-s are distributed continuously, the probability that $\Delta_{ij}(X_k)b - \Delta_{ij}(X_{k'})b = 0$ should be zero for any $b \neq 0$. Define the event $C_{ijkk'}$ (for preserving space) as

$$C_{ijkk'} = \{\omega \in \Omega : \text{sgn}[\delta_{ij}(X_k)] \neq \text{sgn}[\delta_{ij}(X_{k'})]\},$$

so that the previous equation becomes

$$\begin{aligned} Q^{TI}(b) = 2 \ P[C_{ijkk'}] \cdot E\left\{\text{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))\beta] \cdot \right. \\ \left. \cdot \text{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))b] | C_{ijkk'}\right\}. \end{aligned} \tag{1.53}$$

Then the discrepancy between Q^{TI} evaluated at the true value and any other $b \neq \beta$ on the unit circle becomes

$$
\begin{aligned}
Q^{TI}(\beta) - Q^{TI}(b) = 2\ P[C_{ijkk'}]\cdot \qquad (1.54)\\
\cdot E\left\{1 - \operatorname{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))\beta] \cdot \operatorname{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))b] | C_{ijkk'}\right\} = \\
= 2\ P[C_{ijkk'}]\cdot \\
\cdot 2\ E\left[\mathbb{1}[\operatorname{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))\beta] \neq \operatorname{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))b]] | C_{ijkk'}\right] = \\
= 4\ P[C_{ijkk'}]\cdot \\
\cdot P\left[\operatorname{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))\beta] \neq \operatorname{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))b] |\, C_{ijkk'}\right].
\end{aligned}
$$

Note that this is the mathematical representation of the intuition we discussed for the identification lemma, except for X_k and $X_{k'}$ are now random variables.

It is tempting to assume that we proved the (pointwise part of the) identification lemma, because it follows directly that, $Q^{TI}(\beta) - Q^{TI}(b)$ would be minimized at $b = \beta$, while taking the value zero. However, if the sign of the screening values switches with probability zero, or the $\Delta_{ij}(X_k)$ values do not span every direction in $\mathbb{R}^p$, $Q^{TI}(b)$ remains uninformative at least for some b-s.

We need to prove that (1.54) is positive under our standing assumptions whenever $b \neq \beta$. That is, we need to have that the measure of the events $\operatorname{sgn}[\delta_{ij}(X_k)] \neq \operatorname{sgn}[\delta_{ij}(X_{k'})]$ and $\operatorname{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))\beta] \neq \operatorname{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))b]$ conditional on $\operatorname{sgn}[\delta_{ij}(X_k)] \neq \operatorname{sgn}[\delta_{ij}(X_{k'})]$ are not zero under the joint distribution of $\{(X_l, A_l)\}_{l=i,j,k,k'}$.

The conditioning event is not zero measure

According to (1.49) and the symmetry given the i.i.d. assumption, we need to show that

$$P[\Delta_{ij}(X_k)\beta > A_j - A_i > \Delta_{ij}(X_{k'})\beta] > 0.$$

Consider for a small enough $\epsilon > 0$

$$\begin{aligned}
&P[\Delta_{ij}(X_k)\beta > A_i - A_j > \Delta_{ij}(X_{k'})\beta] \geq \qquad (1.55)\\
&\geq P[\Delta_{ij}(X_k)\beta > \epsilon/2, |A_i - A_j| < \epsilon/2, \Delta_{ij}(X_{k'})\beta < -\epsilon/2] =\\
&= E[E[\mathbb{1}[\Delta_{ij}(X_k)\beta > \epsilon/2]\mathbb{1}[\Delta_{ij}(X_{k'})\beta < \epsilon/2]\mathbb{1}[|A_i - A_j| < \epsilon/2]|X_i, X_j]]\\
&\geq E[E[\mathbb{1}[\Delta_{ij}(X_k)\beta > \epsilon/2]\mathbb{1}[\Delta_{ij}(X_{k'})\beta < -\epsilon/2]|X_i, X_j]\cdot\\
&\cdot P[A_j - A_i \in [-\epsilon/2, \epsilon/2]|X_i, X_j]].
\end{aligned}$$

By Assumption 5, the probability of $A_j - A_i \in [-\epsilon/2, \epsilon/2]$ jointly is non-zero conditional on any X_i, X_j realization. Therefore after considering Assumption 5, the remaining statement to prove is that the conditions in the indicator functions are true with non-zero probability (jointly).

Conditional on X_i, X_j the $\Delta_{ij}(X_k)$ and $\Delta_{ij}(X_{k'})$ are independent, and by Claim 6, both $\Delta_{ij}(X_k)$ and $\Delta_{ij}(X_{k'})$ have the ϵ-ball around zero on their support. Together these two statements imply that the joint event $\Delta_{ij}(X_k) > \epsilon/2$ and $\Delta_{ij}(X_{k'}) < -\epsilon/2$ happens with strictly positive probability. Now we established that the probability of $C_{ijkk'}$ being true conditional on X_i and X_j is bigger than zero. We can integrate up with respect to the observables to get that the unconditional probability is also larger than zero.

Possible directions of $\Delta_{ij}(X_k)$

Take any $b \in \bar{B}_\epsilon$. Define $\bar{b}$ as a unit vector in $\mathbb{R}^p$ that is perpendicular to $(\beta + b)/2$. For a $b \in \bar{B}_\epsilon$, the set of possible p-vectors $v \in \mathbb{R}^p$ (the realizations of $\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'})$) for which $\text{sgn}(v\beta) \neq \text{sgn}(vb)$ is an intersection of two open half-spaces by construction, therefore it is a pair of open cones. These cones cannot be empty, because the $\bar{b}$ is in them, so it must contain an open ball of vectors. This means that it is sufficient for point-identification to require that any such pair of cones (corresponding to a pair of spherical wedges with a volume) should have positive probability under the distribution of the observables after conditioning on $C_{ijkk'}$. All the vectors in one cone have positive dot-product with β, and we call this the positive cone (corresponding to the positive wedge), while the other cone is the negative cone.

Now we return to our construction of the previous paragraph, and assume that for some small $\varepsilon > 0$

$$|A_j - A_i| < \varepsilon, \tag{1.56}$$

$$\Delta_{ij}(X_k)\beta > \varepsilon, \tag{1.57}$$

$$\Delta_{ij}(X_{k'})\beta < -\varepsilon \tag{1.58}$$

at the same time. We know that this is a sufficient condition for the $C_{ijkk'}$ event to be true, and that it has also non-zero probability. Without loss of generality for this part of the proof, also assume that the unit ball is on the support of $\Delta_{ij}(X_k)$ (see Claim 6, and consider that we can always rescale the inequalities above). Now we only need that the pair of cones defined by the wedges contain some $\Delta_{ij}(X_k)$ on the positive wedge satisfying (1.57) and $\Delta_{ij}(X'_k)$ on the negative wedge satisfying (1.58),

while $||\Delta_{ij}(X_k)|| = ||\Delta_{ij}(X_{k'})|| = 1$. Due to the cones being open, the inequalities strict and the dot-product continuous, if such one point exists on the support, there is a small ball around that point that has a strictly positive probability, while still satisfying all the constraints we need. This is because the wedges are symmetric around the origo and they are also convex, so the direction of $\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'})$ must be in the positive wedge.

Such a $\Delta_{ij}(X_k), \Delta_{ij}(X_{k'})$ duo always exists, because the wedges have a volume. Define t and $-t$ as the unit vector that represents the direction of the orthogonal projection of β on the hyperplane perpendicular to b. Assume t is on the boundary of the positive wedge. These vectors are such that

$$t\beta = \sin(\phi) > 0, -t\beta = -\sin(\phi) < 0,$$

where ϕ is the angle closed by b and β. Again, since the cones are convex, and the dot-product is continuous, there must be uncountably many vectors v in the positive wedge such that $0 < v\beta < \sin(\phi)$ and $0 > -v\beta > \sin(\phi)$, with $-v$ being in the negative wedge by definition. It follows that $\varepsilon = \sin(\phi)/2$ is a suitably small number to choose. Define v' as the unit vector on the sphere that also agrees with the direction of the orthogonal projection of β on the hyperplane that halves the wedges. Now we can choose anything for $\Delta_{ij}(X_k)$ on the unit sphere that is between t and v' (a weighted average with strictly positive weights, but scaled to unity). Similarly, $\Delta_{ij}(X_{k'})$ can be a weighted average of $-t$ and $-v'$ (with positive weights) scaled to have length one.

Since $P[C_{ijkk'}]$ and the conditional probability term are both strictly positive,

(1.54) is also positive if a normalized $b \neq \beta_0$.

PART 2: Uniform result

The argument so far gave identification in the pointwise sense, that $Q^{TI}(b)$ is uniquely maximized by β (among the unit vectors).

Next we need to prove the continuity of $Q^{TI}(\beta) - Q^{TI}(b)$ in b. For this take a b_n sequence in $\mathbb{R}^p$ that converges to b. The indicator functions at b_n converge to the indicator function at b in the measure defined by the joint distribution of the $(X_l, A_l)_{l \in i,j,k,k'}$ vectors. For this we need again that the joint distribution of these variables is continuous with respect to the Lebesgue measure. Since the indicator at b is measurable, and is dominated by the constant 1, the dominated convergence theorem gives $\lim Q^{TI}(\beta) - Q^{TI}(b_n) = Q^{TI}(\beta) - Q^{TI}(b)$, so $Q^{TI}(b)$ is continuous (sequentially as it maps from $\mathbb{R}^p$ to $\mathbb{R}$ as Euclidean spaces, which is exactly what we need later).

$\bar{B}_\epsilon$ is compact in $\mathbb{R}^p$ as long as $p < \infty$, so by Weierstrass' theorem there exists a $b^*(\epsilon) \in \bar{B}_\epsilon$ such that

$$Q^{TI}(\beta) - Q^{TI}(b^*(\epsilon)) = \inf_{b \in \bar{B}_\epsilon} [Q^{TI}(\beta) - Q^{TI}(b)]. \tag{1.59}$$

Our pointwise result from earlier says that for all $b \in \bar{B}_\epsilon$

$$Q^{TI}(\beta) - Q^{TI}(b) > 0,$$

which gives the desired statement with

$$0 < \delta = Q^{TI}(\beta) - Q^{TI}(b^*(\epsilon)) \leq Q^{TI}(\beta) - Q^{TI}(b) \ \forall b \in \bar{B}_\epsilon. \tag{1.60}$$

□

1.9.4 Fixed effects identification

Let Assumptions 1-7 hold. Then

$$a_j - a_i = E[\Delta_{ij}(X^*)\beta | \delta_{ij}(X^*) = 0, A_i = a_i, A_j = a_j].$$

Proof. Take the case when for an $x_k \in Supp(X_k)$ and $X_i = x_i, X_j = x_j, A_i = a_i, A_j = a_j$, while

$$0 = \delta_{ij}(x_k).$$

This latter equation is equivalent to

$$\begin{aligned} 0 = \delta_{ij}(x_k) = E[D_{ik} - D_{jk} | X_k = x_k, X_i = x_i, X_j = x_j, A_i = a_i, A_j = a_j] = \quad (1.61) \\ = E[F[w'_{ik}\beta + a_i + A_k] - F[w'_{jk}\beta + a_j + A_k]] \Leftrightarrow \\ \Leftrightarrow w'_{ik}\beta + a_i = w'_{jk}\beta + a_j, \end{aligned}$$

where

$$w_{ik} = W_{ik}|_{X_i = x_i, X_k = x_k},$$

the value of W_{ik} after conditioning.

Note that this is not a probabilistic statement. This gives that in general

$$\Delta_{ij}(X_k) - A_i - A_j|_{\delta_{ij}(X_k)=0} = 0 a.s., \quad (1.62)$$

a degenerate 0 random variable, which implies the statement. □

1.9.5 Consistency of the infeasible estimator

Under Assumption 1-8, $\tilde{\beta}^{TI}$ consistently estimates β.

Proof. We follow Sherman (1994) here. Take the normalized statistic

$$\tilde{Q}_n^{TI}(b) - Q^{TI}(b).$$

Then because (the symmetrized) kernel is uniformly bounded by 4, the normalized statistic is a zero-mean U-process of order 4, and the possible kernels (even after symmetrization) belong to a Euclidean class. Then according to Corollary 7 of Sherman (1994) from page 12,

$$\sup_{b \in \bar{B}_\epsilon} |\tilde{Q}_n^{TI}(b) - Q^{TI}(b)| = O_p(1/\sqrt{n}).$$

This means that we have that $\tilde{Q}_n^{TI}$ converges uniformly to Q^{TI}.

The reason why the kernel is Euclidean is that

1. Symmetrization adds up the 'unsymmetrized' kernel finitely many times, and the Euclidean property is closed under (finite) addition.

2. The 'unsymmetrized' kernel is a product of a function that consists of 2 constant segments and $\text{sgn}(\Delta_{ij}(X_k) - \Delta_{ij}(X_k'))b$. The first term of the product is clearly Euclidean, so if the second part is Euclidean, the product is Euclidean as well, since the Euclidean property is closed under (finite) multiplication.

3. The function family of functions

$$\{\text{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X_k'))b] | b \in \mathbb{R}^p : ||b|| = 1\}$$

is Euclidean since the functions $g_b : (\mathbb{R}^p)^4 \to \mathbb{R}$

$$g_b(X_i, X_j, X_k, X_{k'}) = (\Delta_{ij}(X_k) - \Delta_{ij}(X'_k))b$$

for $b \in \mathbb{R}^p$ form a finite dimensional vector space, and so

(a) The family $\{\text{sgn}[(\Delta_{ij}(X_k) - \Delta_{ij}(X'_k))b]\}$ is Euclidean,

(b) the subset of a Euclidean collection is Euclidean.

Here we used numerous well-known lemmas from Nolan and Pollard (1987).

The statement of the lemma follows from the identification result (lemma 3) and uniform convergence by the arguments in Pakes and Pollard (1989) [17]. The reasoning is based on a triangle inequality and can be summarized as

$$\begin{aligned} P[\tilde{\beta}^{TI} \in \bar{B}_\epsilon] &\leq P[\sup_{b \in \bar{B}_\epsilon}(\tilde{Q}_n^{TI}(b) - \tilde{Q}_n^{TI}(\beta)) > 0] \\ &\leq P[\sup_{b \in \bar{B}_\epsilon}(Q(b) - Q^{TI}(\beta)) + O_p(1/\sqrt{n}) > 0] = \\ &= P[-\delta(\epsilon) + O_p(1/\sqrt{n}) > 0] \leq P[O_p(1/\sqrt{n}) > 0] \to 0 \end{aligned} \tag{1.63}$$

□

1.9.6 Consistency of the feasible estimator

1.9.6.1 Effect of the first stage

Under Assumptions 1-9,

$$\lim_{n \to \infty} \sup_{||b||=1} E[|\hat{Q}_n^{TI}(b) - \tilde{Q}_n^{TI}(b)|] = 0. \tag{1.64}$$

[17] In fact, this is enough for $\sqrt{N}$ consistency.

Proof. First estimate the discrepancy using the triangle inequality, then the boundedness of the sign function and random sampling. (Just as before, from the definition of the objective function $i \neq k, k'; j \neq k, k'$.)

$$E|\hat{Q}_n^{TI}(b) - \tilde{Q}_n^{TI}(b)| = \binom{n}{2}^{-1}\binom{n-2}{2}^{-1} E\left|\sum_{i<j,k<k'} \operatorname{sgn}\left[(\Delta_{ij}(X_k) - \Delta_{ij}(X_{k'}))b\right] \cdot \right.$$
$$\left. \cdot \left[\left(\operatorname{sgn}[\hat{\delta}_{ij}(X_k)] - \operatorname{sgn}[\delta_{ij}(X_k)]\right) - \left(\operatorname{sgn}[\hat{\delta}_{ij}(X_{k'})] - \operatorname{sgn}[\delta_{ij}(X_{k'})]\right)\right]\right| \leq$$
$$\leq E\left|\left(\operatorname{sgn}[\hat{\delta}_{ij}(X_k)] - \operatorname{sgn}[\delta_{ij}(X_k)]\right) - \left(\operatorname{sgn}[\hat{\delta}_{ij}(X_{k'})] - \operatorname{sgn}[\delta_{ij}(X_{k'})]\right)\right| \leq$$
$$\leq 2E\left|\operatorname{sgn}[\hat{\delta}_{ij}(X_k)] - \operatorname{sgn}[\delta_{ij}(X_k)]\right| \tag{1.65}$$

So an upper-bound of the L^1 distance is independent of b, and it is of the order of the L^1 distance of the screening value and its kernel estimate.

We know the kernel estimate for $\delta_{ij}(X_k)$ is uniformly consistent[18] under our standing assumptions from Hansen (2008) Theorem 1 on page 729, since it is really a difference of two simple kernel estimator. Since the difference in two sign functions is uniformly bounded, by the dominated convergence theorem this can be strengthened to L^1 convergence. □

1.9.6.2 Consistency proposition

Under Assumptions 1-9, plim $\left(\hat{\beta}^{TI}\right) = \beta$.

Proof. The proof of convergence in Lemma 5 and Lemma 6 proves uniform convergence for $\hat{Q}_n^{TI}(b)$ after using the triangle inequality

$$\sup_b |\hat{Q}_n^{TI}(b) - Q^{TI}(b)| \leq \sup_b |\tilde{Q}_n^{TI}(b) - Q^{TI}(b)| + \sup_b |\hat{Q}_n^{TI}(b) - \tilde{Q}_n^{TI}(b)|.$$

[18] over X_k

Further steps could include the again the standard arguments from Newey and McFadden (1994) using the identification lemma (Lemma 3) and uniform convergence. However, one can directly follow Pakes and Pollard (1989) as well, just as we did to finish the proof of Lemma 5. □

1.9.7 Convergence rate of the simplified estimator

1.9.7.1 Rate of mistake probability

Given Assumptions 1-9 and that f (the pdf corresponding to F) is bounded away from zero and infinity, then for $|\delta_{ij}| > 0$ (which happens almost surely)

$$E[|sgn\hat{\delta}_{ij}(x) - sgn\delta_{ij}(x)||X_{i,j}, A_{i,j}] = P[|sgn\hat{\delta}_{ij}(x) - sgn\delta_{ij}(x)| > 0|X_{i,j}, A_{i,j}] = \\ = O(\exp(-n\sigma_n^p)),$$

however, it is only true that

$$E[|sgn\hat{\delta}_{ij}(x) - sgn\delta_{ij}(x)|] = O(\sqrt{n\sigma_n^p}^{-1})$$

Proof. As before, to preserve space, the conditioning $E[Z|X_i, X_j, A_i, A_j]$ is abbreviated as $E[Z|X_{i,j}, A_{i,j}]$. Also denote the kernel weights corresponding to $D_{il} - D_{jl}$ when calculating screening as $s_n(X_l, x)$, where the n subscript is sometimes omitted. Similarly, if it is not causing any confusion, $\delta_{ij}(x)$ will be written as δ_{ij} only. Note that for $a, b \in \mathbb{R}\backslash\{0\}$

$$|\text{sgn}[a] - \text{sgn}[b]| > 0 \Rightarrow |a - b| > |a|.$$

Now we consider this fact in the pointwise case.

$$P[|\text{sgn}[\delta_{ij}(x)] - \text{sgn}[\hat{\delta}_{ij}(x)]| > 0|X_{i,j}, A_{i,j}] \leq P[|\hat{\delta}_{ij}(x) - \delta_{ij}(x)| > |\delta_{ij}(x)||X_{i,j}, A_{i,j}]$$
$$\leq P[|\hat{\delta}_{ij}(x) - E[\hat{\delta}_{ij}(x)||X_{i,j}, A_{i,j}]| + |E[\hat{\delta}_{ij}(x)|X_{i,j}, A_{i,j}] - \delta_{ij}(x)| > |\delta_{ij}(x)||X_{i,j}, A_{i,j}]$$
$$\leq P[|\hat{\delta}_{ij}(x) - E[\hat{\delta}_{ij}(x)|X_{i,j}, A_{i,j}]| + E[|\hat{\delta}_{ij}(x) - \delta_{ij}(x)||X_{i,j}, A_{i,j}] > |\delta_{ij}(x)||X_{i,j}, A_{i,j}].$$

We need the random sampling assumption and that

$$-2 \leq s(x, X_l)[D_{il} - D_{jl}] \leq 2 \tag{1.66}$$
$$s(x, X_l)[D_{il} - D_{jl}] \perp s(x, X_{l'})[D_{il'} - D_{jl'}]|X_{i,j}, A_{i,j} \tag{1.67}$$

Since if for any three (real) numbers a, b, c it is true that $a > c/2, b > c/2 \Rightarrow a+b > c$,

$$P[|\text{sgn}[\delta_{ij}(x)] - \text{sgn}[\hat{\delta}_{ij}(x)]| > 0|X_{i,j}, A_{i,j}] \leq \tag{1.68}$$
$$\leq P[|\hat{\delta}_{ij}(x) - E[\hat{\delta}_{ij}(x)|X_{i,j}, A_{i,j}]| > |\delta_{ij}/2||X_{i,j}, A_{i,j}]+$$
$$+ P[E[|\delta_{ij}(x) - \hat{\delta}_{ij}(x)||X_{i,j}, A_{i,j}] \geq |\delta_{ij}/2||X_{i,j}, A_{i,j}]$$

Then from (6)-(7),

$$P[|\hat{\delta}_{ij}(x) - E[\hat{\delta}_{ij}(x)|X, A]| > \epsilon|X, A] = \exp\left[-1/2\epsilon^2 \cdot \left(\sum_l \frac{s_n(x, X_l)^2}{(\sum_v s_n(x, X_v))^2}\right)^{-1}\right].$$

by McDiarmid's bounded differences inequality after conditioning on the entire X vector. This is because the largest difference in the concentration inequality $c_l = s_n(x, X_l)/\sum_l s_n(x, X_l) \cdot 2$. We also know that

$$\sum c_l^2 \leq \left(\sum_l s_n(X_l, x)\right)^{-1},$$

so

$$P[|\hat{\delta}_{ij}(x) - E[\hat{\delta}_{ij}(x)|X, A]| > \epsilon|X, A] \leq \exp\left[-1/2\epsilon^2 \cdot \sum_l s_n(X_l, x)\right]. \quad (1.69)$$

After applying the dominated convergence theorem, since we condition on $X_{i,j}, A_{i,j}$, there is a small enough ϵ such that

$$\begin{aligned} &P[|\text{sgn}[\delta_{ij}(x)] - \text{sgn}[\hat{\delta}_{ij}(x)]| > 0||X_{i,j}, A_{i,j}] = \quad (1.70)\\ &E[P[|\hat{\delta}_{ij}(x) - E[\hat{\delta}_{ij}(x)|X, A]| > |\delta_{ij}(x)/2||X, A]|X_{i,j}, A_{i,j}]+\\ &+ P[E[|\delta_{ij}(x) - \hat{\delta}_{ij}(x)||X_{i,j}, A_{i,j}] \geq |\delta_{ij}/2||X_{i,j}, A_{i,j}] \leq\\ &\leq E\left[\exp\left[-1/2\epsilon^2 \cdot \sum_l s_n(X_l, x)\right]|X_{i,j}, A_{i,j}\right] +\\ &+ P[E[|\delta_{ij}(x) - \hat{\delta}_{ij}(x)||X_{i,j}, A_{i,j}] \geq |\delta_{ij}/2||X_{i,j}, A_{i,j}]. \end{aligned}$$

Now we need to apply Hoeffding's inequality for the $\sum s_n(X_l, x)$ once again to realize that out of the randomness coming from the sum, really only its expectation is important for us (up to some constant).

$$\begin{aligned} &E\left[\exp\left[-1/2\epsilon^2 \cdot \sum_l s_n(X_l, x)\right]|X_{i,j}, A_{i,j}\right] = E\left[\exp\left[-1/2\epsilon^2 \cdot \sum_l s_n(X_l, x)\right]\right]\\ &= \exp\left[-1/2\epsilon^2 n \cdot E[s_n(X_l, x)]\right] \cdot\\ &\cdot E\left\{\exp\left[-1/2\epsilon^2 n \cdot \left[n^{-1}\sum_l s_n(X_l, x) - E[s_n(X_l, x)]\right]\right]\right\}, \quad (1.71) \end{aligned}$$

where if $s_n(x, x) = s_0$,

$$P\left[|s_0^{-1}n^{-1}\sum_l s_n(X_l, x) - E[s_n(X_l, x)]| > t\right] \leq \exp(-2nt^2) \quad (1.72)$$

by Hoeffding's inequality. By the law of total probability we have

$$E\left\{\exp\left[-1/2\epsilon^2 s_0 n\cdot\left[s_0^{-1}n^{-1}\sum_l s_n(X_l,x)-E[s_n(X_l,x)]\right]\right]\right\}\leq \tag{1.73}$$
$$\leq \exp(-2nt^2-1/2\epsilon^2 n s_0)+\exp\left[-1/2\epsilon^2 s_0 nt\right]=O(\exp(-C\epsilon^2 n))$$

for a constant C independent of ϵ, if $0<\epsilon<1$ and $t=\epsilon$.

After denoting $t_n=E[s_n(X_l,x)]$, this gives

$$P[|\mathrm{sgn}[\delta_{ij}(x)]-\mathrm{sgn}[\hat{\delta}_{ij}(x)]|>0||X_{i,j},A_{i,j}]= \tag{1.74}$$
$$=O(\exp(-1/2\epsilon^2 nt_n-C\epsilon^2 n))+$$
$$+P[E[|\delta_{ij}(x)-\hat{\delta}_{ij}(x)||X_{i,j},A_{i,j}]\geq|\delta_{ij}/2||X_{i,j},A_{i,j}].$$

Now we realize that if after conditioning on $A_{i,j}$ and $X_{i,j}$, $\delta_{ij}(x)>0$, the second term from (11) (and so in 13 and in 17) is zero, since the pointwise bias is going to zero. That is, the probability of the discrepancy is being greater than the $\epsilon=\delta_{ij}(x)/4$ constant (after conditioning) is *eventually* zero. Note that the first term is uniformly exponentially decreasing.

All in all, for the pointwise case we got that

$$P[|\mathrm{sgn}[\delta_{ij}(x)]-\mathrm{sgn}[\hat{\delta}_{ij}(x)]|>0||X_{i,j},A_{i,j}]=O(\exp(-1/2\epsilon^2 nt_n-C\epsilon^2 n)). \tag{1.75}$$

The argument in the exponential function is always negative as $\epsilon, s_0, t_n>0$, and since t_n is typically decreasing, the exponent is of order $O(nt_n\epsilon^2)$. If ϵ is $O(1)$, then this part of the error probability is $\exp(-O(nt_n))$. Note that if the kernel is a Parzen-Rosenblatt kernel, $O(t_n)=\sigma_n^p$. This gives the result

$$P[|\mathrm{sgn}[\delta_{ij}(x)]-\mathrm{sgn}[\hat{\delta}_{ij}(x)]|>0|X_{i,j},A_{i,j}]\leq C_1\exp(-O(n\sigma_n^p)).$$

The result above is pointwise in $(X_{i,j}, A_{i,j})$, and says something about the rate of the error *eventually*.

If the $|\delta_{ij}(x)|$ had a positive lower bound (taken over $X_{i,j}, A_{i,j}$), then we could simply integrate to get unconditional probability, which would have the same rate even after integration. However, this is not the case. This last step of the proof uses that the probability of $P[|\delta_{ij}| < \epsilon_n] = O(\epsilon_n)$ by the Lipschitz assumption. By the law of total probability,

$$P[|\text{sgn}[\delta_{ij}(x)] - \text{sgn}[\hat{\delta}_{ij}(x)]| > 0] \leq P[\delta_{ij} < \epsilon_n] + \exp(-O(n\sigma_n^p \epsilon^2)) = \tag{1.76}$$

$$= O(\epsilon_n) + \exp[-O(\epsilon_n^2 n\sigma_n^p)]. \tag{1.77}$$

However, it seems we can choose $\epsilon_n^2 = -n\sigma_n^p$, so that the unconditional probability will be $O\left(\left(\sqrt{n\sigma_n^p}\right)^{-1}\right)$. That is, we cannot reach the $O(n^{-1/2})$ level, according to this argument. □

1.9.7.2 Hoeffding-decomposition

Under Assumption 1-10, if b is in an $o_p(1)$ neighborhood of β,

$$Q_n(b,\tau) - Q_n(\beta,\tau) = g(b) - g(\beta) + O_p(\sqrt{n}^{-1})O_p(||b-\beta||) + O_p(n^{-1}) + d_n$$

where

$$d_n(b,\tau) = \binom{n}{2}^{-1} \sum_{i<j} (\tau_{ij} - \tau_0(Z_i, Z_j))[sgn(\Delta_{ij} b) - sgn(\Delta_{ij}\beta)]).$$

Proof. It is clear from the argument we saw in the section where we proved consistency that

$$Q_n(b,\tau) - Q_n(\beta,\tau) = Q_n(b,\tau_0) - Q_n(\beta,\tau_0) + d_n, \tag{1.78}$$

so we can focus on the Hoeffding-decomposition at the non-random τ_0. Everything will follow Sherman (1993, 1994) very closely. After some algebra we arrive at

$$\begin{aligned} Q_n(b,\tau_0) = g(b) + 2n^{-1}\sum_i (f(Z_i) - g) + \\ + \binom{n}{2}^{-1} \sum_{i<j} u^{ij}(Z_i, Z_j) \end{aligned} \tag{1.79}$$

First we need to see that

$$\sup_{b:||b||=1} \binom{n}{2}^{-1} \sum_{i<j} u^{ij}(Z_i, Z_j) = O_p(n^{-1}). \tag{1.80}$$

For this we use the Corollary 4 on page 11 from Sherman (1994). Besides pointing out that the kernel is uniformly bounded, we need to prove that the set of possible kernel functions are Euclidean. Note that here we can regard τ_0 as a known (non-random) function, just as during the first step of the consistency proof. Using the fact that the kernel then is the same as the MRC estimator's by design, we can see that it is Euclidean from the argument in Sherman (1993) page 11.

Our second task is to show that the second term in (1.79) is $O_p[\sqrt{n}^{-1}](b-\beta)$. For this we need that for any b,

$$\sqrt{n}^{-1} \sum_i E[\tau_0(Z_i, Z_j)\text{sgn}(\Delta_{ij} b) - g | Z_i] \tag{1.81}$$

pointwise can be extended around β as

$$f(Z_i;b) = f(Z_i;\beta) + \xi(Z_i;b)(b-\beta) \tag{1.82}$$

$$|\xi(Z_i;b)(b-\beta)| \leq C||b-\beta||, \tag{1.83}$$

where the C constant is independent of b. Following the argument for sufficient conditions in Sherman (1993) to see the existence of this expansion, we can conclude that our smoothness conditions on $f_{X,A}$ are sufficient.

For this consider that

$$\begin{aligned}\sup_b |f(Z_i;b) - f(Z_i;\beta)| \leq \sup_b E[|\tau_0(Z_i,Z_j)(\text{sgn}(\Delta_{ij}b) - \text{sgn}(\Delta_{ij}\beta))||Z_i] \leq \\ \leq E[|\text{sgn}(\Delta_{ij}b) - \text{sgn}(\Delta_{ij}\beta)||Z_i]\end{aligned} \tag{1.84}$$

by the boundedness of τ_0. Now if the b and the β close an α angle, then $|b-\beta| = \sin(\alpha/2)$, and then by a Taylor-expansion of the RHS we have that $O|b-\beta| = O(\alpha)$ if b is in fact a sequence ($\alpha/2 < |b-\beta| \leq e\alpha/2$ as a crude bound). Moreover, if the pdf of the Z_i-s is bounded away from infinity, then there is at most $2c\text{int}(2\pi/\alpha)$ probability mass[19] for which $\text{sgn}\Delta_{ij}b \neq \text{sgn}\Delta_{ij}\beta$. Again, we are interested in the case when $\alpha \to 0$, for which this probability is then $O(\alpha) = O(||b-\beta||)$. This makes $\xi(Z_i;b)$ bounded almost surely, and so by random sampling the triangular CLT applies for given b after demeaning. The same argument can be repeated for

[19]For c being the maximum of the pdf. Here $\text{int}(x)$ is the integer part of the number x.

the $g(b) - g(\beta)$, so

$$n^{-1}\sum_i (f(Z_i;b) - g(b)) - (f(Z_i;\beta) - g(\beta)) = \tag{1.85}$$
$$= n^{-1}\sum_i \xi(Z_i;b)(b-\beta) - (g(b) - g(\beta))$$
$$= O_p(\sqrt{n}^{-1}),$$

because since by the smoothness condition on the distribution functions we have the mean value extension

$$g(b) = g(\beta) + \gamma(b)(b-\beta), \tag{1.86}$$

for some bounded γ function, by the same arguments as in the previous paragraph

$$n^{-1}\sum_i (f(Z_i;b,\tau) - g(b,\tau)) - (f(Z_i;\beta,\tau) - g(\beta,\tau)) = \tag{1.87}$$
$$= n^{-1}\sum_i [\xi(Z_i;b,\tau) - \gamma(b,\tau)](b-\beta)$$
$$= O_p(\sqrt{n}^{-1})(b-\beta),$$

Now if we put together everything,

$$Q_n(b,\tau) - Q_n(\beta,\tau) = g(b) - g(\beta) + O_p(\sqrt{n}^{-1})O_p(||b-\beta||) + O(n^{-1}) + d_n \tag{1.88}$$

Further, note that by Markov's inequality, for some $\epsilon > 0$ constant and r_n positive decreasing sequence

$$P[|d_n| > \epsilon \cdot r_n] \leq E[|d_n|]r_n^{-1}\epsilon^{-1} \leq E[|\hat{\tau}_{ij} - \tau_0(Z_i,Z_j)||\text{sgn}\Delta_{ij}\beta - \text{sgn}\Delta_{ij}b|]r_n^{-1}\epsilon^{-1} \tag{1.89}$$

using the triangle inequality. Also, using that the rate of mistake probability is $O_p(\epsilon_n)$,

$$\sup_{\Delta \in Supp(\Delta_{ij})} E[|\hat{\tau}_{ij} - \tau_0(Z_i, Z_j)||\Delta_{ij} = \Delta] = O_p(\epsilon_n)$$

using the law of total probability after invoking the pointwise result, just as we did before.[20] This gives the upper bound

$$\begin{aligned} P[|d_n| > \epsilon \cdot r_n] &\leq E\left[\sup_{\Delta} E[|\hat{\tau}_{ij} - \tau_0(Z_i, Z_j)||\Delta_{ij} = \Delta]|\text{sgn}\Delta_{ij} b - \text{sgn}\Delta_{ij}\beta|\right] \epsilon^{-1} r_n^{-1} \\ &\leq O_p(\epsilon_n) E[|\text{sgn}\Delta_{ij} b - \text{sgn}\Delta_{ij}\beta|]. \end{aligned} \tag{1.90}$$

As we can see, the rate of the probability of the event that the observable differences give different sign with b and β is still crucial. If we can assume it is $O_p(||b - \beta||)$ as b is in $o_p(1)$ neighborhoods of β, we get that

$$d_n = O_p(\epsilon_n) O_p(||b - \beta||), \tag{1.91}$$

which follows from Assumption 10. □

1.9.7.3 Applying the HPS-lemma

Given Assumptions 1-10, the simplified tetrad inequality estimator $\hat{\beta}$ is approaching the true value β with a non-parametric rate

$$O_p(|\beta - \hat{\beta}|) = O\left(n^{-\frac{1}{2} + \frac{sp}{2}}\right)$$

for $O(\sigma_n) = n^{-s}, 0 < s < 1/p$.

[20] We also use/take into account the Lipschitz-continuity of the cdf F and that the joint pdf $f_{X,A}$ is bounded away from zero.

Proof. I only sketch the proof here heuristically, as it can be found in many papers. We have a sequence of $\hat{\beta}$-s that is approaching the true value β with some rate, and I denote the screening values as $\hat{\tau}$ (the first stage estimates). We will use the Hoeffding-decomposition lemma from the previous paragraphs to get

$$Q_n(\hat{\beta},\hat{\tau}) - Q_n(\beta,\hat{\tau}) = Q(\hat{\beta},\tau_0) - Q(\beta,\tau_0) + O_p(\sqrt{n}^{-1})O_p(||\beta-\hat{\beta}||) + O_p(n^{-1}) + \\ + O_p(\sqrt{n\sigma_n^p}^{-1})O_p(||\beta-\hat{\beta}||). \tag{1.92}$$

First, note that the LHS of this equation is greater than equal to 0, as the $\hat{\beta}$ is maximizing the objective function $Q_n(b,\hat{\tau})$. We need the identification lemma, that the β is uniquely maximizing the $Q(b,\tau_0)$, and that the $Q(b,\tau_0) = g(b,\tau_0)$ can be expanded around β up to the second order (it is continuously differentiable). This is ensured by our assumptions on the joint probability distribution of Z_i. By the necessary strict concavity, there is a $\kappa > 0$ for which

$$Q(b,\tau_0) - Q(\beta,\tau_0) \leq -\kappa||b-\beta||^2,$$

which then gives

$$O_p(||\hat{\beta}-\beta||^2) \leq O(\sqrt{n\sigma_n^p}^{-1})O_p(||\beta-\hat{\beta}||) + O_p(n^{-1}), \tag{1.93}$$

According to the argument in Sherman (1994) (the general method, lemma 1), after completing the square, the rate at which $\hat{\beta}$ approaches β is the minimum of the $\sqrt{n}$ and $\sqrt{n\sigma_n^p}$. □

1.9.8 Estimation of fixed effects

Given a vanishing sequence $r_n \geq \sqrt{n}^{-1}$ *eventually, such that* $\sup_{i,j,k} |\hat{\delta}_{ij}(X_k) - \delta_{ij}(X_k)| = O_p(r_n)$*, and* $||\hat{\beta}-\beta|| = O_p(r_n)$*, define* $\hat{\alpha}_{ij}$ *as above. If Assumption 1-7*

hold,

$$|\hat{\alpha}_{ij} - (a_i - a_j)| = O_p\left(\max\left\{\sigma_n^l, r_n(\sigma_n^l)^{-2}\right\}\right)$$

with zero expectation.

Proof. Denote the kernel weights by $l_n(\hat{\delta}_{ij}(X_k)) = \frac{L_n(\delta_{ij}(X_k))}{\sum_k L_n(X_k)}$, where $L_n(x) = L\left(\frac{x}{\sigma_n^l}\right)$. Note that we condition on $A_i = a_i, A_j = a_j, X_i = x_i, X_j = x_j$ throughout this section. This conditioning will be abbreviated as $|A_{i,j}, X_{i,j}$.

By linearity,

$$\begin{aligned} l_n[\hat{\delta}_{ij}(X_k)]\Delta_{ij}(X_k)\hat{\beta} =& l_n[\delta_{ij}(X_k)]\Delta_{ij}(X_k)\beta + \\ &+ \Delta_{ij}(X_k)(\hat{\beta} - \beta)l_n[\hat{\delta}_{ij}(X_k)] + \\ &+ \Delta_{ij}(X_k)\beta(l_n[\hat{\delta}_{ij}(X_k)] - l_n[\delta_{ij}(X_k)]) \end{aligned} \tag{1.94}$$

Note that the $|\Delta_{ij}(X_k)(\beta)|$ and[21] the $|\hat{\delta}_{ij}(X_k)|$ are bounded from above. Also, because the kernel is assumed to be smooth, there is a mean-value expansion of $L_n(\hat{\delta}_{ij}(X_k))$ around $L_n(\delta_{ij}(X_k))$ so that $|L_n(\hat{\delta}_{ij}(X_k)) - L_n(\delta_{ij}(X_k))| = O_p(\sup|\delta_{ij}(X_k) - \hat{\delta}_{ij}(X_k)|) = O_p(r_n)$. Considering these claims, after applying the Cauchy-Schwartz inequality it is immediate from the definition of r_n that

$$\begin{aligned} |\sum_k \Delta_{ij}(X_k)(\hat{\beta} - \beta)l_n[\hat{\delta}_{ij}(X_k)]| \leq \sup_k |\Delta_{ij}(X_k)(\hat{\beta} - \beta)| \leq \\ \sup_k ||\Delta_{ij}(X_k)|| \cdot ||\hat{\beta} - \beta|| = O_p(r_n). \end{aligned} \tag{1.95}$$

[21] We assumed bounded support as part of compactness on a Euclidean space.

On the other hand,

$$|\sum_k \Delta_{ij}(X_k)\beta(l_n[\hat{\delta}_{ij}(X_k)] - l_n[\delta_{ij}(X_k)])| = O_p\left(r_n(\sigma_n^l)^{-2}\right). \tag{1.96}$$

To see this, we need to do some tedious algebra to get

$$\begin{aligned}
&\sum_k \Delta_{ij}(X_k)\beta(l_n[\hat{\delta}_{ij}(X_k)] - l_n[\delta_{ij}(X_k)]) = \\
&= \frac{\sum_k L_n(\hat{\delta}_{ij}(X_k))\Delta_{ij}(X_k)\beta}{\sum_{k'} L_n(\hat{\delta}_{ij}(X_{k'}))} \frac{\sum_{k'} L_n(\delta_{ij}(X_{k'})) - L_n(\hat{\delta}_{ij}(X_{k'}))}{\sum_{k'} L_n(\delta_{ij}(X_{k'}))} - \\
&- \frac{\sum_k [L_n(\delta_{ij}(X_k)) - L_n(\hat{\delta}_{ij}(X_k))]\Delta_{ij}(X_k)\beta}{\sum_{k'} L_n(\delta_{ij}(X_{k'}))}
\end{aligned} \tag{1.97}$$

According to this,

$$\begin{aligned}
&\left|\sum_k \Delta_{ij}(X_k)\beta(l_n[\hat{\delta}_{ij}(X_k)] - l_n[\delta_{ij}(X_k)])\right| \leq \\
&\leq \left|\frac{\sum_k L_n(\hat{\delta}_{ij}(X_k))\Delta_{ij}(X_k)\beta}{\sum_{k'} L_n(\hat{\delta}_{ij}(X_{k'}))}\right| \cdot \left|\frac{\sum_{k'} L_n(\delta_{ij}(X_{k'})) - L_n(\hat{\delta}_{ij}(X_{k'}))}{\sum_{k'} L_n(\delta_{ij}(X_{k'}))}\right| + \\
&+ \left|\frac{\sum_k [L_n(\delta_{ij}(X_k)) - L_n(\hat{\delta}_{ij}(X_k))]\Delta_{ij}(X_k)\beta}{\sum_{k'} L_n(\delta_{ij}(X_{k'}))}\right|
\end{aligned} \tag{1.98}$$

after using the triangle inequality.

Again, the first term of the product is a (weighted) average of bounded numbers, which means that it is itself bounded by the bound. Call the

$$\sup_{x_k \in Supp(X_k)} \Delta_{ij}(x_k)\beta = \bar{\Delta}.$$

Then after using the triangle inequality again,

$$\left|\frac{\sum_k [L_n(\delta_{ij}(X_k)) - L_n(\hat{\delta}_{ij}(X_k))]\Delta_{ij}(X_k)\beta}{\sum_{k'} L_n(\delta_{ij}(X_{k'}))}\right| \leq \frac{\sum_k |L_n(\delta_{ij}(X_k)) - L_n(\hat{\delta}_{ij}(X_k))|}{\sum_{k'} L_n(\delta_{ij}(X_{k'}))}\bar{\Delta}, \tag{1.99}$$

so this means

$$\left|\sum_k \Delta_{ij}(X_k)\beta(l_n[\hat{\delta}_{ij}(X_k)] - l_n[\delta_{ij}(X_k)])\right| \leq 2\frac{\sum_k |L_n(\delta_{ij}(X_k)) - L_n(\hat{\delta}_{ij}(X_k))|}{\sum_{k'} L_n(\delta_{ij}(X_{k'}))}\bar{\Delta}. \tag{1.100}$$

This means we need to get the rate of $\frac{\sum_k |L_n(\delta_{ij}(X_k))-L_n(\hat{\delta}_{ij}(X_k))|}{\sum_{k'} L_n(\delta_{ij}(X_{k'}))}$. Now by our assumption on the kernel (Lipschitz) we have that

$$(n-2)^{-1}\sum_k |L_n(\delta_{ij}(X_k)) - L_n(\hat{\delta}_{ij}(X_k))| = (\sigma_n^l)^{-1}O_p(r_n), \tag{1.101}$$

while

$$\begin{aligned}(n-2)^{-1}\sum_k L_n(\delta_{ij}(X_k)) &= E[L_n(\delta_{ij}(X_k))|X_{i,j}, A_{i,j}]+ \\ &+ [(n-2)^{-1}\sum_k L_n(\delta_{ij}(X_k)) - E[L_n(\delta_{ij}(X_k))|X_{i,j}, A_{i,j}]].\end{aligned} \tag{1.102}$$

The first term is going to zero with the rate of σ_n^l, because the kernel L_n is of Parzen-Rosenblatt type. Moreover, the second term shows the discrepancy of the average of bounded i.i.d. random variables from their mean, which is a random variable that is $O_p(\sqrt{n}^{-1}) = o_p[\sigma_n^l]$ by assumption (and by for example Hoeffding's inequality of a CLT). After collecting the terms we get that

$$(n-2)^{-1}\left|\sum_k \Delta_{ij}(X_k)\beta(l_n[\hat{\delta}_{ij}(X_k)] - l_n[\delta_{ij}(X_k)])\right| = O_p\left(d_n(\sigma_n^l)^{-2}\right).$$

Now we turn to the analysis of the first term, which actually gives the information we seek. A Taylor-expansion gives

$$F[W_{ik}\beta + A_i + A_k] - F[W_{jk}\beta + A_j + A_k] = f(\xi)[\Delta_{ij}(X_k)\beta + A_i - A_j], \tag{1.103}$$

for some $0 < M^{-1} < f(\xi)^{-1} < m^{-1} < \infty$ by the bilipschitz assumption, which implies

$$\begin{aligned} M^{-1}|F[W_{ik}\beta + A_i + A_k] - F[W_{jk}\beta + A_j + A_k]| \leq \\ \leq |\Delta_{ij}(X_k)\beta - (A_j - A_i)| \leq \\ \leq m^{-1}|F[W_{ik}\beta + A_i + A_k] - F[W_{jk}\beta + A_j + A_k]|, \end{aligned} \tag{1.104}$$

which after taking expectation over A_k using the conditional distribution of the fixed effects we get

$$M^{-1}|\delta_{ij}(X_k)| \leq |\Delta_{ij}(X_k)\beta - (A_j - A_i)| \leq m^{-1}|\delta_{ij}(X_k)|. \tag{1.105}$$

Then from here

$$\left|\sum_k l_n[\delta_{ij}(X_k)][\Delta_{ij}(X_k)\beta - (A_j - A_i)]\right| \leq m^{-1}\sum_k l_n[\delta_{ij}(X_k)]|\delta_{ij}(X_k)| \tag{1.106}$$

using the triangle inequality. Using again the original notation of the kernel weights, clearly,

$$\frac{\sum_k L_n[\delta_{ij}(X_k)]|\delta_{ij}(X_k)|}{\sum_k L_n[\delta_{ij}(X_k)]} \to 0, \tag{1.107}$$

because eventually any fix $\delta_{ij}(X_k) >> 0$ will have $L_n()$ weight that is converging to zero, while we know that the denominator is $O_p(n\sigma_n^l)$ from the kernel density estimator literature under usual kernel rates.

In the coming equation I will omit the conditioning on $X_{i,j}, A_{i,j}$, but it is

understood throughout. The numerator can be characterized as

$$E\left[\sum_k L_n[\delta_{ij}(X_k)]|\delta_{ij}(X_k)|\right] = n\int L(u)(\sigma_n^l)^2|u|f_{\delta_{ij}}(u\sigma_n^l)du = \tag{1.108}$$
$$= n(\sigma_n^l)^2\int L(u)|u|f_{\delta_{ij}}(0)du + n(\sigma_n^l)^3\int L(u)|u|f'_{\delta_{ij}}(\xi)du =$$
$$= O(n(\sigma_n^l)^2)$$

and

$$V\left[\sum_k L_n[\delta_{ij}(X_k)]|\delta_{ij}(X_k)|\right] = n\int L^2(u)(\sigma_n^l)^3|u|^2 f_{\delta_{ij}}(u\sigma_n^l)du = \tag{1.109}$$
$$= n(\sigma_n^l)3\int L^2(u)|u|^2 f_{\delta_{ij}}(0)du + n(\sigma_n^l)^4\int L(u)|u|f'_{\delta_{ij}}(\xi)du =$$
$$= O((\sigma_n^l)^3 n)$$

after respective Taylor-expansions (only up to the first order) around zero. All in all, we can replicate the results of the theory of the Nadaraya-Watson estimator. To summarize, after using Slutsky's theorem, there are such sequences for which

$$\sqrt{n(\sigma_n^l)^{-1}}\left(\frac{\sum_k L_n[\delta_{ij}(X_k)]|\delta_{ij}(X_k)|}{\sum_k L_n[\delta_{ij}(X_k)]} - O(\sigma_n^l)\right) = \tag{1.110}$$
$$= \sqrt{n(\sigma_n^l)^{-1}}\left(\frac{(n\sigma_n^l)^{-1}\sum_k L_n[\delta_{ij}(X_k)]|\delta_{ij}(X_k)|}{(n\sigma_n^l)^{-1}\sum_k L_n[\delta_{ij}(X_k)]} - O(\sigma_n^l)\right) = O_p(1).$$

Again, even if we know the true values, the bottleneck is the bias, which is the same rate as the bandwidth according to this argument.

All in all, conditional on $A_i = a_i, A_j = a_j, X_i = x_i, X_j = x_j$ for some unknown a_i, a_j, we have that after invoking the triangle inequality,

$$|\hat{\alpha}_{ij} - \alpha_{ij}| = O_p\left(\max\left\{r_n, r_n(\sigma_n^l)^{-2}, \sigma_n^l, \sqrt{n(\sigma_n^l)^{-1}}^{-1}\right\}\right) \tag{1.111}$$

Note that this is only a pointwise result. If r_n is converging to zero at most at a $\sqrt{n}$ rate, the first and the last sequences are eventually lower then the middle ones. $\square$

Chapter 2

Nonparametric identification of distance functions in network formation models with fixed effects

2.1 Introduction

In this paper we provide non-parametric identification arguments for the set of models where the outcome is binary and the possible unobserved heterogeneity terms enter additively in the model. The main goal of the paper is to show the identification and estimation argument corresponding to a non-parametric version of the empirical strategy for the network formation model in Toth (2018). In this model, the indicator (D_{ij}) of a connection between node i and j can be written as

$$D_{ij} = \mathbb{1}[w(X_i, X_j) + A_i + A_j \geq \epsilon_{ij}], \tag{2.1}$$

where the random vector X_i is observable, w is an unknown distance function, and A_i are the fixed effect terms that can be arbitrarily correlated with the observables, while the pair-specific noise ϵ_{ij} is exogenous, conditional on the fixed effects. As shown in Toth (2018), identification of the parametric linear index model rests on a version of double-differencing. The main goal of this paper is achieved by restating the inequality-type identification argument into a conditioning type estimator. The main tool for estimation is a series expansion of w, as due to the linearity of the

approximation, we can follow the information we get from double-differencing in a simple way.

As an additional benefit of the series estimators, the generalization results in an estimator that does not require the numerical maximization of an objective function. This is an important point, as we would like to increase the number of dimensions of observables as the sample size grows, and the available numerical optimization methods would become less reliable.[1] Using standard results from Newey (1997) and Belloni et al. (2015), we also characterize the asymptotic behavior of the resulting estimators. As a drawback, the guaranteed achievable rate is at most half of the optimal non-parametric rate in Stone (1982). While this is not unexpected, it is especially serious in our particular application, network formation, where increasing the sample size will give less information due to the sparsity present in these questions.[2]

We conclude that while using parametric strategies and driving the number of dimensions of the parameter space to infinity is an intuitive approach to non-parametric estimation, but it is not without caveats. In particular, the researcher needs to ensure identification *up to scale* on the limit, separately from the finite dimensional identification arguments. Nonparametric identification results are

[1] Although for Han's estimator we have some results in Wang (2007).

[2] We understand sparsity as a limiting property of the network as the number of nodes grows to infinity. The data generating process does NOT reflect this property, and our asymptotics are not designed to predict how the same estimator would work compared to the true values in a network with a million sample size. Our asymptotics are informative about how good of an approximation the estimate is for the parameters that generated the network given its size in our snapshot.

seldom of this type (see Matzkin 2006 for overview), which means that we cannot blindly use the series approach.

As it has been shown in Toth (2018), double-differencing is a much more difficult problem than first differencing. To present the main idea more clearly, we will also give the main argument for the classical problems of Han (1987) and Manski (1987), when the dimension of unobserved heterogeneity is zero and 1, respectively. The simple cases again highlight the duality between the classical inequality type reasoning from the above papers and their conditional counterparts defined here, which could be viewed as a variant of Ahn (1995), Blundell and Powell (2004) or Ahn, Ichimura and Powell (2004) to name a few. In this section we restate some results from this literature in disguise. Besides the above results, the panel literature mainly continued on the inequality track (for example Cavanagh and Sherman (1998), Abrevaya (2000), Abrevaya and Shin (2011), Khan and Tamer (2010)). Related to this, a similar question is considered by Matzkin (1989), but our results are more general, even for the simple cases. In addition, for example Froelich (2006) also uses the same series type approach to estimated structural functions in binary outcome models locally, when there is no unobserved heterogeneity. While their results are not directly applicable, Fan et al. (2017) deals with maximum rank correlation type estimators with increasing number of observables.[3]

This paper shows how the differencing (screening) approach from Toth (2018) is applicable to the nonparametric case, and it produces an identification result

[3]They assume away identification problems, and relatedly, their observables are independent across dimensions.

for the network formation model. A parallel work of Guo (2017) also establishes identification results of multiple, sometimes more general version of our model up to various normalizations. His approach can also be understood in the framework of Toth (2018), and a similar strategy is considered in Toth (2018b). However, without an additional *identifying* assumption on the support of observables, generalizing the identification argument would further increase computational complexity and decrease the rate of convergence by another degree. This limits the practical applicability of the use of more convoluted identification arguments, and makes simple approaches important. In addition to this, this paper is the first to provide a consistent estimator for the nonparametric case.

In the first section we give identification and estimation results for the generalized regression models with and without fixed effects. We only propose the estimators, as their consistency and rate results could be derived analogously after reading the analysis of our main topic. In the second section, first we ask under what normalizations the differencing information from screening can identify our parameter of interest, the distance function. Then we define an estimator and research its properties. Section 4 concludes after brief Monte Carlo simulations in section 3.

On notation: since we are ran out of letters early on, and we want to point out parallels throughout, we use the same letter to denote different objects in different sections. However, this is always mentioned when introducing the object et each section. Moreover, $||.||$ is the Euclidean-norm throughout. For a vector X_i, if it is a variables that observed through multiple time period ($T = 0, 1$), then

$X_i = [X_{i0}\ X_{i1}]$. Generically, the ith element of a vector v is denoted as v^i.

2.2 Simplified problem and main idea

Consider the generalized regression model

$$Y_{it} = D[F[v(X_{it}), A_i, \epsilon_{it}]], \tag{2.2}$$

where D is a weakly, while F is a strictly increasing scalar function, X_{it} is a p-vector of observables, T is the number of time periods we observe the same set of individuals, A_i is a fixed effect unobservable of dimension 1, and ϵ_{it} is a one-dimensional continuous random variable. All three functions D, F and v or the distribution of ϵ_{it} are unknown. For the sake of simplicity, we will only look at $T = 2$. Our parameter of interest in this section is $v(.)$.

The main identifying assumptions on the structure of the model are summarized in Assumption 11. The assumptions are fairly standard, and proposed to match Han's and Manski's seminal paper.[4]

Assumption 11 (Simplified model.)

Given the data generating process in (2.2)*,*

1. *Exogeneity: ϵ_{it} is i.i.d. distributed with a cdf F_ϵ and independent of X_i conditional on the fixed effects*

2. *(L_1-)Invertibility:*

[4]These are not necessary conditions. Also note that there is a redundancy between 2b) and 2a), given Assumption 11, WLOG we can assume D is a step function.

(a) F is strictly increasing

(b) D is a step function or it is strictly increasing on a non-zero measure interval of the support of $F[v(X_{it}), A_i, \epsilon_{it}]$

(c) the conditional support $Supp(F[v(X_{it}), A_i, \epsilon_{it}]|X_i, A_i) =]\infty^-, \infty^+[$ a.s.

(d) the conditional pdf of $F[v(X_{it}), A_i, \epsilon_{it}]|_{X_i=x_i, A_i=a_i}$ is bounded away from zero on every compact interval of the support for every (a_i, x_i).

3. Smoothness: F and v are continuously differentiable

4. Locally non-convex v: for any $x \in \mathbb{R}^p$ and $\epsilon > 0$,

$$\sup_{y \in B_\epsilon(x)} v(y) > \inf_{y \in B_\epsilon(x)} v(y).$$

The identification argument of Manski (1987) rests on homogeneity assumptions on unobservables through time. To our knowledge, every marginal rank correlation type estimator can be summarized by some zero-set, where the differences in the conditional expectation of a given function of outcomes are exactly zero. This aligns with the intuition provided by screening from Toth (2018), and this differencing is the equivalent of screening from Toth (2018) in these models. Manski's paper exploits time-homogeneity, so we need to focus on the conditional expectation

$$E[Y_{i1} - Y_{i0}|X_i] = 0 \tag{2.3}$$

This results in a screening lemma of the form

Lemma 9 (Screening lemma for $T = 2$.)

Take the simplified model. Given Assumption 11,

$$E[Y_{i1} - Y_{i0}|X_i] = 0 \Leftrightarrow v(x_{i1}) - v(x_{i0}) = 0 \tag{2.4}$$

The screening lemma gives us *equivalence relationships* under which for any x_i, x_j we can determine if $v(x_{i1}) - v(x_{i0}) = 0$, just by looking at screening values. As the sample size goes to infinity, the information we are given is the level curves of $v(x)$ on the (common) support of the X_i-s. For $x, y \in Supp(X_i)$

$$x \sim y \Leftrightarrow v(x) = v(y) \Leftrightarrow S(x, y) = 0, \tag{2.5}$$

where the screening value[5] $S(x, y)$ is defined as

$$S(x, y) = E[Y_{i1} - Y_{i0}|X_i = (x, y)] \text{ if } T = 2. \tag{2.6}$$

The question is, given this information, what do we know about $v(.)$. This problem is very well-known for everybody in economics, as it is the simplest version of a utility representation problem. Except in our case, it already follows that a "utility function" (v) representing a total order on the support of the observables exists. So by the same argument as in graduate micro 101, after assuming continuity for v, we can identify it up to a bijective transformation on the support. This means that from Assumption 11 we can get a Matzkin-type (after Matzkin (2006)) identification result summarized in the following corollary.

[5]It is assumed to be identified at this point.

Corollary 1

Given Assumption 11, the ratios of partial derivatives at any $x_0 \in Int(Supp(X_i))$

$$\frac{\frac{\partial v(x_0)}{\partial x_0^q}}{\frac{\partial v(x_0)}{\partial x_0^r}}$$

are identified for $1 < q, r < p$ *integers, whenever the denominator is not zero and the support of* X^r *and* X^q *are both convex around* x_0*.*

The proof is quite straightforward, as the ratio of partials is the derivative of the identified level curve. One could also introduce a local derivative estimator to estimate this ratio pointwise, but since this result is not the main point of the paper, we do not analyze this route further.

Instead, we point out that our conclusion is consistent with the structure of the classical inequality-type estimators. In Manski's objective function, if we observe the sign of the difference of outcome variables, we get a probabilistic information about the linear order of the X_i vectors corresponding to the ordering defined by $v()$. To put it in another way, after taking conditional expectations,

$$\text{sgn}\{E[Y_{i1} - Y_{i0}|X_{i0} = x, X_{i1} = y]\} = \text{sgn}\{v(y) - v(x)\}.$$

In the binary case the object

$$E[\text{sgn}(Y_{i1} - Y_{i0})|X_{i0} = x, X_{i1} = y] = E[Y_{i1} - Y_{i0}|X_{i0} = x, X_{i1} = y],$$

is very simple to calculate.[6] Manski's estimator expresses the same information content with the asymmetric binary relation that corresponds to $\sim$ above, except

[6]This is not going to be the case when we need double-differencing in our main model.

for one detail. The equivalence relationship from the screening lemmas above does not give a natural ordering, while the inequality type arguments (based on the asymmetric binary relation) do.[7] As it turns out, our question requires identification up to scale, for which we need additional assumptions.

2.2.1 Main idea

The idea is that the information from screening is given as a difference, so a linear approximation can preserve it, since linear maps commute with differencing. For example, assume that there is a sequence of functions $f_m : \mathbb{R}^p \to \mathbb{R}$ such that there exists a sequence of coefficients $\{c_m\}$ for which

$$v(x) = \sum_{m=0} c_m f_m(x). \tag{2.7}$$

Then

$$0 = v(x) - v(y) \Leftrightarrow 0 = \sum_{m=0} c_m (f_m(x) - f_m(y)). \tag{2.8}$$

Such linear approximations are convenient, as we can often incorporate other assumptions as restrictions on the coefficients, so that v corresponds to a unique sequence of coefficients. Another nice feature is that if the regulatory conditions are met regarding the $\{f_m\}$ sequence from Newey (1997), the resulting consistent estimator is weighted OLS, after normalizing one of the coefficients to 1.

The caveat is that we can only gain information from the variation when $x \sim y$, when the screening values are the same. Moreover, there will be some m

[7] Hence the reason why the information from the inequality-type screening lemma, like in Toth (2018) would identify v up to a strictly positive monotone transformation, but the conditioning type approaches only to a strictly monotone transformation.

indices for which $f_m(x) - f_m(y) = 0$, for all x, y. In particular, the regression we use for identification reads in its full form as

$$\mathbb{1}_{S(x,y)=0}[f_q(x) - f_q(y)] = \mathbb{1}_{S(x,y)=0} \sum_{m=q+1} \frac{c_m}{c_q}(f_m(x) - f_m(y)), \tag{2.9}$$

where q is the first integer for which $f_m(x) - f_m(y) \neq 0$, and we need to assume that $c_q \neq 0$.[8] These considerations also show that the information we use from these semiparametric/nonparametric arguments always result in identification up-to-scale, and the number of additional normalizations is q. This is the best-case scenario, when we have already found the suitable sequence of functions $\{f_m\}$, for which the conditions in Newey (1997) are satisfied.

Take the sequence of functions $\{f_m\}_{m=0}^{\infty}$, and arrange them in a way such that for the first q of them we have $f_m(x) - f_m(y) = 0$ for (x, y) a.e. Define

$$\Delta_{N,M} = \begin{bmatrix} f_{q+1}(x_1) - f_{q+1}(y_1) & \dots & f_m(x_1) - f_m(y_1) & \dots & f_M(x_1) - f_M(y_1) \\ \dots & & & & \\ f_{q+1}(x_i) - f_{q+1}(y_i) & \dots & f_m(x_i) - f_m(y_i) & \dots & f_M(x_i) - f_M(y_i) \\ \dots & & & & \\ f_{q+1}(x_N) - f_{q+1}(y_N) & \dots & f_m(x_N) - f_m(y_N) & \dots & f_M(x_N) - f_M(y_N) \end{bmatrix} \tag{2.10}$$

Assumption 12 (Sufficient variation and approximation.)

For the simplified model,

1. *Existence of linear approximation: $v \in \mathcal{V}$. There is a known sequence of functions $\{f_m : \mathbb{R}^p \to \mathbb{R}\}_m$ such that for every $u \in \mathcal{V}$ there exists a unique*

[8] Naturally, for identification purposes we only need one such c_m, and if we do not have any, we identified that $v(x) = 0$.

corresponding sequence of coefficients $\{c_m^u\}_m \in \mathcal{C}$*, for which*

$$\sup_{x \in Supp(X_i)} \left| u(x) - \sum_{m=0}^{M} c_m^u f_m(x) \right| \leq O(r_N) = o(1)$$

where $\forall u \in \mathcal{V}$*,* $r_N^u \leq O(r_N) = o(1)$*.*

2. *Approximation rate:* $M \to \infty$ *as* $N \to \infty$

3. *Sufficient variation:*

   (a) *The eigenvalues of* $E[\Delta'_{N,M}\Delta_{N,M}]$ *are bounded away for any* M

   (b) X_i *is a* $\mathbb{R}^p$*-valued continuous random variable with compact support that is not a proper subspace of* $\mathbb{R}^p$

Regarding the support assumption of our observables, if the researcher encounters discrete variables with small support, the identification result can be used to identify the structural functions after fixing the discrete variables.

Define

$$\hat{w}_i^M = K_N[\hat{S}^M(x_{i0}, x_{i1})], \tag{2.11}$$

$$\hat{S}^M(x_{i0}, x_{i1}) = \frac{\sum_{l \neq k} \kappa_N(x_{k0} - x_{i0}, x_{k1} - x_{i1})(y_{k0} - y_{k1})}{\sum_{l \neq k} \kappa_N(x_{k0} - x_{i0}, x_{k1} - x_{i1})}, \tag{2.12}$$

$$\Delta_i^q = f_q(x_{i0}) - f_q(x_{i1}) \tag{2.13}$$

$$\Delta_i = [\Delta_i^{q+1}\ \Delta_i^{q+2}\ ...\ \Delta_i^M], \tag{2.14}$$

where κ_N is the corresponding kernel function with bandwidth σ_N. Given $M = M(N)$ (only in subscripts), our estimator can be summarized as

$$\hat{\beta}_M^M = \min_{b_M} N^{-1} \sum_i \hat{w}_i^M \left[f_q(x_{i0}) - f_q(x_{i1}) - \Delta_i b_M \right]^2. \tag{2.15}$$

Let us define $\hat{W}^M$ the $N \times N$ matrix that has $\hat{w}_i^M$ as its diagonal element in the ith row. In this case $\Delta_{N,M}$ is going to be a $N \times M$ matrix, consisting of the Δ_i vectors stacked in order. Similarly, Δ^q is the stacked version of the LHS variables. Then

$$\hat{\beta}_M^M = [\Delta'_{N,M}\hat{W}^M\Delta_{N,M}]^{-1}\Delta'_{N,M}\hat{W}^M\Delta^q. \tag{2.16}$$

The presence of the fixed effects needs us to require that the inverse of $E[u(v(x_i), A_i)|v(x_i)]$ is Lipschitz. This is automatically satisfied if we have a semi-additive index model

$$Y_{it} = \mathbb{1}[v(x_{it}) + A_i > \epsilon_{it}]$$

with an ϵ_{it} that has unbounded support and strictly increasing cdf. In this case the same kind of argument as included at the section 2.3.2 can be made to derive consistency and rate of convergence results for the simplified models.

2.2.2 Application and identification caveat

In economics, researchers like to interpret their identifying assumptions and normalizations. For this reason we will look at the result for Manski's model when the chosen approximation is the Taylor-series. This already restricts the set where v must belong ($\mathcal{V}$) to the set of analytic functions.[9] Moreover, the numerical stability of this series may pose a problem in individual applications, because the polynomials are not orthogonalized. On the other hand, normalization and identification assumptions on the coefficients are easy to interpret.

[9]If the original $\mathcal{V}$ is a Hilbert space of some continuous functions that is the superset of analytic functions, we can get an analogue identification result that is sufficient for estimation purposes.

For this subsection, given $z \in \mathbb{R}^p$, define the sequence of functions and coefficients as

$$f_\alpha(x) = (x - z)^\alpha, \tag{2.17}$$

$$c_\alpha = \frac{\alpha!}{\partial^\alpha v(z)}, \tag{2.18}$$

where α is a multi-index, and we are going to use the multi-index notation throughout this paper. Then if v is an analytic function, we have that

$$S(x, y) = 0 \Rightarrow v(x) - v(y) = \sum_{|\alpha|>0} c_\alpha (f_\alpha(x) - f_\alpha(y)) = 0. \tag{2.19}$$

Here the constant $(0th)$ term of the series cancels out, and so a necessary location normalization we need is that

$$v(z) = 0. \tag{2.20}$$

Moreover, we are inclined to set

$$c_{(1,0,\ldots,0)} = \left.\frac{\partial v(x)}{\partial x^1}\right|_{x=z} = 1, \tag{2.21}$$

which means that we restrict the slope of the function v in the first dimension at the anchor point z to 1. Another nice property of analytic functions is that they are either zero everywhere, or only zero on zero-measure places. This gives us that the remaining terms of the f_α-differences are going to be non-zero almost everywhere.[10]

With the sufficient variation condition the Gram-matrices would be non-singular, however we need more than that. We need to ensure that the point

[10] We can trivially find a point where they are non-zero.

spectrum of the limiting infinite matrix does not contain zero. To state it differently, Corollary 1 states that the $v(x)$ is only identified up to a positive *monotone transformation*. If it worked, our procedure would imply that it is identified up to *scale*. There are in fact infinitely many $\{\tilde{c}_n\}$ sequences for which (2.19) is satisfied. For example, take the function $\tilde{v}(x) = 0.5[\exp(2v(x)) - 1]$, which is still going to be analytic, it has the same first derivative if we normalize $v(z) = 0$ as $v(z)$, but the ratio of the coefficient sequence is clearly different otherwise from the true sequence. This is because the assumption above is not enough, we have identification failure in general. Denote the set of possible coefficients that correspond to the normalized version of $v \in \mathcal{V}$-s by $\mathcal{C}$.

Lemma 10

Given Assumption 12, and corresponding regulatory conditions in Newey (1997) on the rate of M, in the simplified model it is necessary and sufficient for identification and the consistency of $\hat{\beta}$ that the program

$$\min_{\{c_n\}\in\mathcal{C}} E\left\{\left[\sum c_n(f_n(x) - f_n(y))\right]^2 | S(x,y) = 0\right\}$$

has a unique solution. In particular, if v is not identified up to scale in the simplified problem, $\lim \Delta'_{N,M}\Delta_{N,M}$ has zero on its points spectrum, $\lim \hat{\beta}$ is ill-defined.

This means it is necessary to establish identification separately, and find some restriction A on v that lets us identify the function up to scale, so that we know if

$$A \Leftrightarrow \{c_n\} \in \mathcal{C}_A \subsetneq \mathcal{C},$$

then

$$\min_{\{c_n\}\in\mathcal{C}_A} E\left\{\left[\sum c_n(f_n(x)-f_n(y))\right]^2 |S(x,y)=0\right\}$$

has a unique solution. If this was not the case, our procedure may still pick a closest function in the span of the f_m-s for any finite M, but the coefficients would only get infinitesimally close to the identified set.[11]

There are more than one possible restrictions here, besides homogeneity and concavity from Matzkin (1989). A good choice would be to assume a numeraire or additivity, if the researcher thinks of v as expected utility. The following proposition sums up this application.

Proposition 4

Take the simplified problem with $T=2$, and v analytic. Given Assumptions 11-12, if either

$$v(x) = g(x^2,...,x^p)x^1 + h(x^2,...,x^p)$$

for some unknown functions h, g or

$$v(x) = \sum_{i=1}^{p} g_i(x^i),$$

for some unknown functions g_i, then the true $\{c_n\}$ sequence corresponding to v (and the function itself) is identified up to scale and location.

This means that besides the location and scale normalization in (2.20)-(2.21), we need that $c_\alpha = 0$ for every α either with $\alpha^1 \geq 2$, or with more than one positive

[11] This itself does not give much information, potentially; see the discussion in Khan and Tamer (2007).

entry.

Note that in this paper we talk about the identifying properties of the information rank-correlation type *estimators* give us, not about the identifiability of the *models* themselves. The same caveat is important for the generalization for Han's estimator, but we know that after assuming ϵ_{i0} is uniformly distributed on the $[0, 1]$ interval, the link function itself is identified. This identifying assumption on the other hand is not easy to translate into our framework.

There are many other, possibly better choices for the approximation sequence. In particular, we conjecture the Fourier series (with $\mathcal{V}$ is L^1) and multivariate splines/B-splines ($\mathcal{V}$ is C^p) being possibly more efficient choices. For a small comparison and further reference, please consult Newey (1997) and/or Belloni et al. (2015).

2.3 Nonparametric estimation of distance functions

Now we apply the same idea for our main problem. As mentioned in the introduction, we observe the adjacency matrix D of a network, the $[i, j]$th element of which tells us if node i and j in the network are connected or not. We also observe the observable characteristics of every node ($X_i \in \mathbb{R}^p$), but when the node is born into the network, it is endowed with an unobservable characteristic A_i as well. The vector (X_i, A_i) is an independent draw from the same distribution for every i, but the elements can depend on each other. Given the unknown function $w : \mathbb{R}^{2p} \to \mathbb{R}$ and pair-specific unobservable ϵ_{ij}, we assume that the links between

node i and j are formed according to

$$D_{ij} = \mathbb{1}[w(X_i, X_j) + A_i + A_j \geq \epsilon_{ij}]. \tag{2.22}$$

The intuition behind $w(X_i, X_j)$ is that it represents some kind of distance. Therefore, in the literature it is often assumed that it has the basic properties of a geometric distance. Assumption 13 fills this role for us.

Assumption 13 (Distance function.)

The function w has the following properties:

1. *Zero property: $w(x, x) = 0$ for any $x \in \mathbb{R}^p$,*
2. *Symmetry: $w(x, y) = w(y, x)$ for any $x, y \in \mathbb{R}^p$,*
3. *Smoothness: w is continuously differentiable.*

Another important group of identifying assumptions is the conditions needed for our screening lemma.

Assumption 14

In the main model, we assume

1. *Exogeneity: ϵ_{ij} is i.i.d. distributed with cdf F_ϵ and independent of X_i conditional on the fixed effects*
2. *Invertibility: F_ϵ is strictly increasing on the whole real line,*
3. *Common support: $Supp(A_i|X_i = x) = Supp(A_i|X_i = x')$ for every $x, x' \in Supp(X_i)$.*

4. *Smoothness: F_ϵ is continuously differentiable*

5. *Locally non-convex w: for any $x, z \in \mathbb{R}^p$ and $\epsilon > 0$, $\sup_{y \in B_\epsilon(x)} w(y, z) > \inf_{y \in B_\epsilon(x)} w(y, z)$.*

As it is discussed in Toth (2018), these are only sufficient assumptions for the screening lemma below. In particular, we can weaken the independence assumption on the disturbance terms, and we only need to require that there is *some* common conditional support of the fixed effects.

For the main model, define the screening value

$$S_{ij}(x) = E[D_{ik} - D_{jk} | X_i = x_i, A_i = a_i, X_j = x_j, A_j = a_j, X_k = x]. \tag{2.23}$$

As it has been established in Toth (2018), the appropriate screening lemma is as follows.

Lemma 11 (Screening lemma for the main model.)
In the network formation model, given Assumption 14, for any $x, x' \in Int(Supp(X_i))$

$$S_{ij}(x) = 0 \ \wedge \ S_{ij}(x') = 0 \Leftrightarrow w(x_i, x) - w(x_j, x) = a_j - a_i = w(x_i, x') - w(x_j, x'). \tag{2.24}$$

Proof. This is a corollary of the screening lemma in Toth (2018). □

In our main model we need double-differences to cancel out the terms of unobserved heterogeneity:

$$S_{ij}(x) = 0 \ \wedge \ S_{ij}(x') = 0 \Leftrightarrow (w(x_i, x) - w(x_j, x)) - (w(x_i, x') - w(x_j, x')) = 0.$$

As emphasized by Toth (2018) and for example Charbonneau (2017), in this two-dimensional fixed effect model we cannot 'easily' calculate the double-differences of the conditional expectations, and a straightforward generalization of Manski's approach does not exist. Moreover, the time homogeneity assumptions are very different in nature, as this case requires additivity in the index, so that the comparisons during screening work properly. Besides additivity, the common support condition can also be considered a time homogeneity assumption. It ensures that in some ball around zero, whatever value the difference $w(x_i, x) - w(x_j, x)$ takes, there will be some realization values of A_i, A_j such that $0 \in Supp(S_{ij}(x))$.

Lastly, the information about $w(.,.)$ is more restricted than what we would have under simple double differencing $(v(x_i) - v(x_j) - v(x_k) + v(x_l))$, as we cannot vary the inputs as freely. This is going to be an important detail, because as opposed to the finite linear case in Toth (2018), we cannot achieve non-parametric identification without varying all four legs of the tetrad (i, j, k, l).

Similarly to the previous case, the information is provided as a difference. We can look at it in multiple ways. First, using Corollary 1 we could think of it as identifying

$$\frac{\frac{\partial(w(x_i,x)-w(x_j,x))}{\partial x^1}}{\frac{\partial(w(x_i,x)-w(x_j,x))}{\partial x^q}} \tag{2.25}$$

on the support. Equivalently, *after fixing* x_i, x_j, we trace out the level curves of the difference function $\Delta_{ij} : \mathbb{R}^p \to \mathbb{R}$

$$\Delta_{ij}(x) = w(x_i, x) - w(x_j, x),$$

or identify the $\Delta_{ij}(x)$ up to a strictly monotone transformation. The key problem is that this transformation can be different for every (x_i, x_j) couple.

Denote

$$\Delta^q_{ijkl} = f_q(x_i, x_k) - f_q(x_j, x_k) - f_q(x_i, x_l) + f_q(x_j, x_l) \tag{2.26}$$

$$\Delta_{ijkl} = [\Delta^{q+1}_{ijkl}\ \Delta^{q+2}_{ijkl}\ ...\ \Delta^{M}_{ijkl}], \tag{2.27}$$

and let $\Delta_{N,M}$ be the stacked version of the copies Δ_{ijkl}-s (one row for every tetrad). To apply the main idea for generalization, we need the following assumption.

Assumption 15 (Sufficient variation and approximation.)

For the main model,

1. *Existence of linear approximation: $w \in \mathcal{W}$. There is a known sequence of functions $\{f_m : \mathbb{R}^q \to \mathbb{R}\}_m$ such that for every $u \in \mathcal{W}$ there exists a unique corresponding sequence of coefficients $\{c^u_m\}_m \in \mathcal{C}$, for which*

$$\sup_{x \in Supp(X_i)} \left| u(x) - \sum_{m=0}^{M} c^u_m f_m(x) \right| < O(r_M) = o(1)$$

where $\forall u \in \mathcal{W}$.

2. *Approximation rate: $M = M(N) \to \infty$ as $N \to \infty$*

3. *Sufficient variation:*

   (a) *The eigenvalues of $E[\Delta'_{N,M}\Delta_{N,M}]$ are bounded away from zero over M*

   (b) *X_i is a $\mathbb{R}^p$-valued continuous random variable that is not a proper subspace of $\mathbb{R}^p$*

2.3.1 Identification

Now we conduct an identification analysis.

2.3.1.1 Information from the zero-property

Our main framework makes it convenient to incorporate the information from Assumption 13 using already introduced objects. How do we produce a distance function? Take any continuously differentiable function $\tilde{w} : \mathbb{R}^{2p} \to \mathbb{R}$, and define

$$w(x, y) = \tilde{w}(x, y) + \tilde{w}(y, x) - \tilde{w}(y, y) - \tilde{w}(x, x). \tag{2.28}$$

In particular, $w(x, y)$ can always be written as

$$w(x_i, x_j) = \frac{w(x_i, x_j) + w(x_j, x_i) - w(x_j, x_j) - w(x_i, x_i)}{2} = \frac{\Delta^w_{ijji}}{2}. \tag{2.29}$$

This means that the distance functions are fixed points of the above operator, which has a double-difference structure (up to a constant multiple). This shows how well the structure from Assumption 13 fits to our source of information, which is a restriction on double-differences.

Denote

$$\Delta^m_{ijkl} = f_m(x_i, x_k) - f_m(x_j, x_k) - f_m(x_i, x_l) + f_m(x_j, x_l)$$

In particular, after using the linear approximation for the RHS, we get that

$$w(x_i, x_j) = \frac{1}{2} \sum c_m \Delta^m_{ijji}. \tag{2.30}$$

Lemma 12

Assume f_m are continuous. If $m < q$, that is, the term f_m cancels after double-differencing Δ^m_{ijkl} almost everywhere as a function of the quadruple (x_i, x_j, x_k, x_l), then $\Delta^m_{ijji} = 0$.

This means that due to the distance function assumption, our framework can *potentially* identify the relevant members of the true sequence $\{c_m\}$ up to scale, if the sufficient variation condition in Assumption 15 is satisfied. However, given the identification caveat in the case of the simple models, it is still a question when we can apply our method. For this section, denote

$$m_{ik} = \frac{\partial m(w(x_i, x_k), x_i)}{\partial w(x_i, x_k)},$$
$$w_{jk^1} = \frac{\partial w(x_j, x_k)}{\partial x^1_k}.$$

As expressed earlier, from the information provided by screening, we can identify

$$\frac{\frac{\partial (w(x_i,x) - w(x_j,x))}{\partial x^1}}{\frac{\partial (w(x_i,x) - w(x_j,x))}{\partial x^q}}, \tag{2.31}$$

for fixed x_i, x_j and at any x and $1 < q \leq p$ if the denominator is not vanishing.[12] A simple way of proceeding is through the following small lemma.

Lemma 13

Given Assumption 13,

$$\left.\frac{\partial w(x, y)}{\partial x^q}\right|_{x=y} = 0 \ \forall x.$$

[12]Note that the bijection m must be continuously differentiable, since the smoothness assumption on the possible set of w-s in Assumption 13.

That is, the first order partial derivatives of w when its two arguments equal are zero.

Proof. By the zero property, for any q

$$w(x+dx^q, x+dx^q) = 0 = w(x,x) + \left.\frac{\partial w(x,y)}{\partial x^q}\right|_{x=y} dx^q + \left.\frac{\partial w(x,y)}{\partial y^q}\right|_{x=y} dx^q, \tag{2.32}$$

which gives by symmetry that

$$0 = \left[\left.\frac{\partial w(x,y)}{\partial x^q}\right|_{x=y} + \left.\frac{\partial w(x,y)}{\partial y^q}\right|_{x=y}\right] dx^q = 2\left.\frac{\partial w(x,y)}{\partial x^q}\right|_{x=y} dx^q, \tag{2.33}$$

which gives the required conclusion. □

Since the object is identified in (2.31), even when $x = x_j$, together with the previous lemma from the same information we also identify

$$\left.\frac{\frac{\partial(w(x_i,x)-w(x_j,x))}{\partial x^1}}{\frac{\partial(w(x_i,x)-w(x_j,x))}{\partial x^q}}\right|_{x=x_j} = \frac{\frac{\partial w(x_i,x_j)}{\partial x_j^1}}{\frac{\partial w(x_i,x_j)}{\partial x_j^q}}, \tag{2.34}$$

which means we can back out $w(x_i, x)$ up to an injective continuous transformation $m((.), x_i)$, and if d is in the class of admissible distance functions that rationalizes the data, it can be written as

$$d(x_i, x_j) = m(w(x_i, x_j), x_i) \ \forall x_i, x_j$$

Then we know that since the ratio in (2.31) is identified,

$$\frac{m_{ik}w_{ik^1} - m_{jk}w_{jk^1}}{m_{ik}w_{ik^q} - m_{jk}w_{jk^q}} = \frac{w_{ik^1} - w_{jk^1}}{w_{ik^q} - w_{jk^q}}, \tag{2.35}$$

which gives after some algebra

$$(m_{ik} - m_{jk})(w_{ik^q} w_{jk^1} - w_{ik^1} w_{jk^q}) = 0. \tag{2.36}$$

On the one hand, if

$$\frac{w_{ik^1}}{w_{ik^q}} = \frac{w_{jk^1}}{w_{jk^q}}, \tag{2.37}$$

it means that the level sets of the distance functions from x_i and x_j have parallel level curves at x_k. We can assume this case away, as we have already identified the shape of these functions, and this is not an interesting case for our applications. However, also note that by ruling out the equality in 2.37, we assume that there are at least two dimensions affecting the distance measure[13].

On the other hand, if the shape of the level curves change for a non-zero measure convex set $G \subset Supp(X_i)^3$, if $(x_i, x_j, x_k) \in G$ we have an unknown function g for which

$$m_{ik} = m_{jk} = g(x_k). \tag{2.38}$$

This would imply after integrating up from

$$d_{i^1 k} = m_{ik} w_{i^1 k} = g(x_k) w_{i^1 k} \tag{2.39}$$

that for a constant C,

$$d(x_i, x_k) = g(x_k) w(x_i, x_k) + C. \tag{2.40}$$

[13]This itself is a testable assumption using the approach with screening values from Toth (2018).

Then we once again use symmetry to deduce

$$d(x_i, x_k) = \frac{d(x_i, x_k) + d(x_k, x_i)}{2} = \frac{g(x_k) + g(x_i)}{2} w(x_i, x_k), \tag{2.41}$$

$$d(x_j, x_k) = \frac{d(x_j, x_k) + d(x_k, x_j)}{2} = \frac{g(x_k) + g(x_j)}{2} w(x_j, x_k), \tag{2.42}$$

$$\Rightarrow g(x_k) = \frac{g(x_k) + g(x_i)}{2} = \frac{g(x_k) + g(x_j)}{2} \Leftrightarrow g(x_i) = g(x_j) = g(x_k) = constant. \tag{2.43}$$

This shows that there is a neighborhood on which $w(x_i, x_j)$ is identified up to scale, and if the set of (x_i, x_j, x_k) for which

$$\frac{w_{ik^1}}{w_{ik^q}} = \frac{w_{jk^1}}{w_{jk^q}}$$

is zero-measure, then it is identified up to scale on the support.[14] It turns out that this is the case when we can assume another metric property, that

$$w(x, y) = 0 \Rightarrow x = y, \tag{2.44}$$

so the distance is strictly positive between two non-identical points. From the discussion above, the following result follows

Proposition 5

Given the main model with Assumptions 13-14, if for $x_i, x_j, x_k \in Supp(X_i)$ there are $q, s \in \mathbb{N}, 1 \leq q < s \leq p$ such that

$$\frac{w_{ik^s}}{w_{ik^q}} \neq \frac{w_{jk^s}}{w_{jk^q}},$$

[14]Note that for our application below this does not matter, as local identification of w is equivalent to global identification. This is exactly what we are paying for by restricting w to be analytic.

then w is identified up to an affine transformation by its level curves on the neighborhood of (x_i, x_k).

Moreover, if the set of such (x_i, x_j, x_k) are measure 1, sufficiently, if $w(x, y) = 0 \Leftrightarrow x = y$, then w is identified up to scale by the information from the screening lemma.

2.3.1.2 Using the Taylor-expansion

In this subsection we expand w around $(z_1, z_2) \in Int(Supp(X_i)^2)$, after assuming that it is analytic. However, first to show the expansion we use:

$$w(x, y) = \sum_{|\alpha|>0} \frac{([x\ y] - [z_1\ z_2])^\alpha}{\alpha!} (\partial^\alpha w)(z_1, z_2).$$

Here α is a multi-index, and we think of it as a row vector of length $2p$. It is comprised of two vectors (multi-indices) stacked horizontally:

$$\alpha = [\alpha_1\ \alpha_2],$$

and we also define the $'$ operation as

$$\alpha' = [\alpha_2\ \alpha_1].$$

After first differencing every term that was not indexed by i or j will cancel out:

$$w(x_i, x_k) - w(x_j, x_k) = \sum_{|\alpha|>0} \frac{([x_i\ x_k] - [z_1\ z_2])^\alpha - ([x_j\ x_k] - [z_1\ z_2])^\alpha}{\alpha!} (\partial^\alpha w)(z_1, z_2).$$

To be more specific,

$$f_\alpha(x_i, x_k) - f_\alpha(x_j, x_k) = 0 \Leftrightarrow \alpha_1 = 0$$

Now we can take double-differences. From the remaining terms those will cancel out, whichever are functions of x_i and/or x_j only.

$$\Delta_{ijkl} = \sum_{|\alpha|>0} (\partial^\alpha w)(z_1, z_2) \cdot \frac{([x_i\ x_k] - [z_1\ z_2])^\alpha - ([x_j\ x_k] - [z_1\ z_2])^\alpha - ([x_i\ x_l] - [z_1\ z_2])^\alpha + ([x_j\ x_l] - [z_1\ z_2])^\alpha}{\alpha!} \tag{2.45}$$

For example, this means that double-differencing only leaves information about the coefficients that are in the off-diagonal matrix of the Hessian up to order two. This will generalize to the higher-order terms as well. To make this explicit,

$$\Delta_{ijkl}^{f_\alpha} = 0 \Leftrightarrow \alpha_1 = [0, ..., 0] \text{ OR } \alpha_2 = [0, ..., 0]. \tag{2.46}$$

We can conclude that due to double-differencing, we only get information about the coefficients from the expansion for which $|\alpha_1| \cdot |\alpha_2| > 0$. Luckily, as argued above, due to Assumption 13, we do not need any other ones. To apply our main idea for this expansion, we need the normalization

$$c_{[1,0,..,0;1,0,...,0]} = \frac{\partial^2 w(z_1, z_2)}{\partial z_1^1 \partial z_2^1} = 1. \tag{2.47}$$

Then we need to restrict our coefficients to induce symmetry with the following:

$$c_\alpha = c_{\alpha'}\ \forall \alpha. \tag{2.48}$$

2.3.1.3 Alternative assumptions

While the model is non-parametrically identified under our assumptions, the identifying variation may be very small. For example, in the argument above we relied heavily on the calculation of the slope of the level curves around the endpoints of the $\overrightarrow{x_i x_k}$ vector. Similarly to the simplified models, we can come up with assumptions that make the model stronger, and also have nice interpretation.

Superposition/Homogeneity

If we assume that the distance function only depends on the difference vector, that is

$$w(x, y) = t(x - y),$$

for some function $t : \mathbb{R}^p \to \mathbb{R}^+$, then we do not need Assumption 13 for identification, although we will need location and scale normalization, so we will impose the zero-property implicitly.

The reason is that the screening lemma for the main model does not need the special assumptions on w. Moreover, with the above mentioned location normalization $t(0) = 0$, we can still arrive to the conclusion that we can identify the distance function $t(x - y)|_{x=x_i} = \bar{t}(x_i - y)$ up to bijective smooth transformation. However, this time the transformation can be regarded as the same for every x_i, given the homogeneity/superposition assumption.

Additivity

If we assume additivity, we can identify the differenced function

$$w(x_i, x_k) - w(x_j, x_k)$$

up to scale right after invoking Corollary 1. For example using the zero-property, we can get $w(x_i, x_j)$ up to scale after plugging in $x_k = x_j$. Symmetry, or in case of other location normalizations, even the zero-property can be relaxed. The compelling side of the additivity assumption is that we can allow for almost everywhere differentiable functions, which would not rule out the absolute value function as distance function. Moreover, with additivity, we would not need to vary all four legs of the tetrad observables (x_i, x_j, x_k, x_l), and so a consistent simplified estimator like in Toth (2018) would exist.

2.3.2 Estimation

In this section we define our estimator for the main model and analyze its asymptotic behavior. Partly as a reminder, we define

$$\hat{w}_{ijkl} = K_N[\hat{S}_{ij}(x_k)]K_N[\hat{S}_{ij}(x_l)], \tag{2.49}$$

$$\hat{S}_{ij}(x_k) = \frac{\sum_l \kappa_N(x_k - x_l)(D_{il} - D_{jl})}{\sum_l \kappa_N(x_k - x_l)}, \tag{2.50}$$

$$\Delta^q_{ijkl} = f_q(x_i, x_k) - f_q(x_j, x_k) - f_q(x_i, x_l) + f_q(x_j, x_l) \tag{2.51}$$

$$\Delta_{ijkl} = [\Delta^{q+1}_{ijkl}\ \Delta^{q+2}_{ijkl}\ ...\ \Delta^{M}_{ijkl}], \tag{2.52}$$

where κ_N is the corresponding kernel function with bandwidth σ_N. Given $M = M(N)$ (only in subscripts), our estimator can be summarized as

$$\hat{\beta}_M = \min_{b_M} \binom{N}{2}^{-1} \binom{N-2}{2}^{-1} \sum_{i<j} \sum_{k<l} \hat{w}_{ijkl} \left[\Delta^q_{ijkl} - \Delta_{ijkl} b_M\right]^2. \tag{2.53}$$

Call $N_4 = \binom{N}{2}^{-1}\binom{N-2}{2}$. Let us define $\hat{W}$ the $N_4 \times N_4$ matrix that has $\hat{w}_{ijkl}$ as its diagonal elements in lexicographic order of the index. In this case $\Delta_{N_4,M}$ is going

to be a $N_4 \times M$ matrix, consisting of the Δ_{ijkl} vectors stacked in the same order, and Δ^q is the stacked version of the LHS variables in 2.53. Then

$$\hat{\beta}_M = [\Delta'_{N_4,M}\hat{W}\Delta_{N_4,M}]^{-1}\Delta'_{N_4,M}\hat{W}\Delta^q. \tag{2.54}$$

We make the following assumptions about the estimation parameters and the sample generating process

Assumption 16 (Estimation assumptions.)

The data for our main model satisfies the following conditions.

1. *Sampling: the vector (X_i, A_i) is sampled independently from a common distribution*

2. *Smoothness1: The pdf of (X_i, A_i), f_{X_i,A_i} and the strictly increasing F_ϵ is twice continuously differentiable. Moreover, the pdf of characteristics is bounded away from zero and infinity, hence the joint support is compact*

3. *Smoothness2: $w(.,.)$ has uniformly bounded partial derivatives on (the bounded) $Supp(X_i)^2$*

4. *Bandwidth assumptions: $h_N, \sigma_N \to 0$ and $\sigma_N = O\left(\left(\frac{\log N}{N}\right)^{\frac{1}{p+4}}\right)$*

5. *The kernel K_N is the uniform kernel.*

Denote

$$\xi_M = \sup_{(x,y)} ||f(x,y)|| = O(M^{-k}),$$

and the rate r as

$$\sup_{(x,y)\in Supp(X_i)^2} \inf_{\omega\in span(\{f_m\}_1^M)} ||w(x,y) - \omega(x,y)|| = M^{-r}.$$

Both r and ξ_M are discussed objects in Newey (1997) and Belloni et al. (2015) for multiple choices of series. Also define

$$\hat{w}(x_i, x_j) = \sum_{m=1}^{M} \beta_m \Delta^m_{ijji}$$

Proposition 6

Given Assumption 13-16, and that $r \leq 1/p, k \geq 1/2$. *If* $\xi_M^2 \frac{\log M}{N h_N^{-2}} = o(1)$, *the estimator* $\hat{\beta}$ *is consistent, and any small enough* $\epsilon > 0$

$$||w(x) - \hat{w}(x)|| = o_p\left(N^{-\frac{r}{2(k+r)}+\epsilon}\right) \tag{2.55}$$

The proof of this proposition is included in the Appendix.

We conclude that the rate of the estimator is not going to reach the optimal rate of the simple non-parametric regression, and in every examples for series in Belloni et al. (2015) or Newey (1997), the rate will actually be slower than half of the optimal rate. Note in the parametric specification of Toth (2018), the optimal rate was achievable by the tetrad inequality estimator.

2.4 Monte Carlo simulation

For this Monte Carlo simulation we assume that the true distance function is

$$w(x_i, x_j) = (x_i^1 - x_j^1)^2(x_i^2 + x_j^2) + (\exp(x_i^2) - \exp(x_j^2))^2 \tag{2.56}$$

The data generating process is described by

$$X_i^{1,2} \sim U[-2,2] \text{or} TN[-2,2,-1/0,2] \tag{2.57}$$

$$A_i = 0.25X_i^1 - 0.5|X_i^2| \cdot Z_i,\ Z_i \sim N_{[0,1]}[1,2] \tag{2.58}$$

$$U_{ij} \sim \text{logistic}[0,1] \tag{2.59}$$

$$D_{ij} = \mathbb{1}[w(x_i,x_j) + A_i + A_j \geq \epsilon_{ij}] \tag{2.60}$$

We estimate the w function using the estimator $\hat{\beta}$ for the coefficients when the Taylor-expansion is used with $z_1 = (0.05, 0.1), z_2 = (-0.05, -0.1)$. The sample sizes are $100, 300$ and 500, and the bandwidth selection follows the rates prescribed in the previous section. In Table 2.1 we calculate a simulated value of the L_2-distance between the estimated $\hat{w}$ and the normalized w (IRMSE column).

$N = 100$	Effective sample size mean	IRMSE mean
$p^M = 2$	73,963	0.463
$p^M = 3$	73,963	0.405
$p^M = 4$	73,963	0.929
$N = 300$	Effective sample size mean	IRMSE mean
$p^M = 2$	2,750,437	0.404
$p^M = 3$	2,750,437	0.288
$p^M = 4$	2,750,437	0.829
$N = 500$	Effective sample size mean	IRMSE mean
$p^M = 2$	14,778,047	0.385
$p^M = 3$	14,778,047	0.251
$p^M = 4$	14,778,047	0.781

Table 2.1: Monte Carlo results for the conditional series estimator. The number of repetitions is 500. We calculated the sample size eventually entering into the least squares regressions in the second column. Then we included the simulated value of the L_2 error for the function (using the uniform density on the support) in the last column. The p^M numbers denote the orders up to which the polynomial series were included in the regression.

Due to the small sample size, the simulation suggests we cannot really include the terms of the fourth order ($p^M = 4$). As discussed above, $p^M = 2$ has only two approximating function, but $p^M = 3$ has 9, and $p^M = 4$ already 23. As we can see, the bias is relatively high, at best it is around 5% of the range of the function. On the other hand, the norms are decreasing at the rates that is expected from the theoretical results in the previous section. It is expected that around $N = 2,500$ one would include the $p^M = 4$ order terms.

2.5 Conclusion

In this paper we non-parametrically identified the distance function in the network formation model of Toth (2018) and Graham (2017). We highlighted the importance of the 'metric type' assumptions on distance function for identification, as they naturally translate into a normalization that is needed for double-differencing type approaches. The argument suggests an estimator that is based on a linear approximating series. We show this estimator to be consistent. This approach also results in a closed form estimator of the coefficients that determine the unknown distance function, which is an advantage, as numerical optimization methods in high dimensions tend to be less reliable for rank correlation estimators.
Since the method described above is flexible, we introduce it on simpler models that include Han's generalized regression and Manski's semi-parametric panel regression models.

2.6 References

Abrevaya, J. and Shin, Y. (2011). Rank estimation of partially linear index models. The Econometrics

Abrevaya, J. (2000). Rank estimation of a generalized fixed-effects regression model, Journal of Econometrics, Volume 95, Issue 1, 2000, pp. 1-23.

Abrevaya, J. (1999). Rank estimation of a transformation model with observed truncation. The Econometrics Journal, 2: 292-305.

Ahn, H., (1995). Nonparametric two-stage estimation of conditional choice probabilities in a binary choice model under uncertainty, Journal of Econometrics, Volume 67, Issue 2, 1995, 337-378,

Ahn, H., Ichimura, H., Powell, J. L. and Ruud, P. A. (2015). Simple Estimators for Invertible Index Models, WP

Ahn, H. and Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. Journal of Econometrics, Volume 58, Issues 1–2, 1993, pp. 3-29.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. Journal of Econometrics, 2015, vol. 186, issue 2, 345-366

Blundell, R. W. and Powell, J.L. (2004). Endogeneity in Semiparametric Binary Response Models. Restud Review of Economic Studies (2004) 71, 655–679

Cavanagh, C. and Sherman, R. P. (1998). Rank estimators for monotonic index models. Journal of Econometrics, 84(2):351–381.

Charbonneau, K. B. (2017), Multiple fixed effects in binary response panel data models. The Econometrics Journal, 20: S1-S13.

Dragomir, S. S. (2015). Reverses of Schwarz inequality in inner product spaces with applications. Math. Nachr., 288: 730-742. doi:10.1002/mana.201300100

Stinchcombe, M. B. and Drukker, D. M. (2014). Regression efficacy and the curse of dimensionality. WP

Fan, Y., Han, F., Li, W., Zhou, X-W. (2017). On Rank Estimators in Increasing Dimensions, WP

Froelich, M. (2006). Non-parametric regression for binary dependent variables. The Econometrics Journal, Vol. 9, No. 3 (2006), pp. 511-540

Graham, B. S. (2015). An econometric model of link formation with degree heterogeneity. Technical Report 20341, National Bureau of Economic Research. Published in Econometrica (2017).

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. Journal of Econometrics, 35(2-3):303–316.

Hansen, B. E. (2008). Uniform Convergence Rates for Kernel Estimation with Dependent Data. Econometric Theory, Vol. 24, No. 3 (Jun., 2008), pp. 726-748

He, X. and Shao, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. The Annals of Statistics, 24(6):2608–2630.

Hoderlein, S. and White, H. (2012,2010) Nonparametric identification in nonseparable panel data models with generalized fixed effects, Journal of Econometrics, Volume 168, Issue 2, 2012, Pages 300-314,

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. The Annals of Mathematical Statistics, 19(3):293–325.

Honore, B. E. and Powell, J. (2005). Pairwise difference estimators for nonlinear models. In Andrews, D.W.K., Stock, J.H. (Eds.) Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg, pages 520–553. Cambridge University Press.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 221–233. Berkeley, CA.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. The Annals of Statistics, 1(5):799–821.

Khan, S. and Tamer, E. (2007). Partial rank estimation of duration models with general forms of censoring. Journal of Econometrics, 136(1):251–280.

Kiefer, J. (1967). On Bahadur's representation of sample quantiles. The Annals of Mathematical Statistics, 38(5):1323–1342.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. Journal of Econometrics, 3(3):205 – 228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. Journal of Econometrics, 27(3):313 – 333.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. Econometrica, 55(2):357 – 362.

Matzkin, Rosa. (1989). A Nonparametric Maximum Rank Correlation Estimator. Cowles Foundation WP.

Matzkin, R., (2007). Nonparametric identification. In: Handbook of Econometrics, 2007, vol. 6B, Chapter 73, Elsevier

Newey, W. (1997). Convergence rates and asymptotic normality for series estimators. Journal of Econometrics, 1997, vol. 79, issue 1, 147-168

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. The Annals of Statistics, 15(2):780–799.

Pakes, A., and Pollard, D. (1989). Simulation and the Asymptotics of Optimization Estimators. Econometrica, 57(5), 1027-1057.

Rudelson, M. (1999). Random Vectors in the Isotropic Position. Journal of Functional Analysis, Volume 164, Issue 1, 60-72,

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. John Wiley and Sons.

Sherman, R. P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. The Annals of Statistics, 22(1):439–459.

van der Vaart, A. and Wellner, J. (1996). Weak Convergence and Empirical Processes. Springer.

Wang, H. (2007). A note on iterative marginal optimization: a simple algorithm for maximum rank correlation estimation. Computational Statistics and Data Analysis, 51(6):2803–2812.

2.7 Appendix

2.7.1 Proof of Lemma 9

Since the proof is nearly identical to the previous case, we proceed faster. Fix x_i, and denote

$$u[v(x_{i0}), a_i] = E[Y_{i0}|X_i = x_i, A_i = a_i]. \tag{2.61}$$

The function u is strictly increasing in its first argument if we hold the second constant. This is proven the same way as previously, except we need to rewrite the equations with an additional a_i in them.[15]

$$v(x) > v(y) \Rightarrow F[v(x), a_i, \epsilon] - F[v(y), a_i, \epsilon] > 0 \ \forall \epsilon, a_i \in Supp((a_i, \epsilon_{it})|_{X_i=x}), \tag{2.62}$$

so taking the step function case as an example, for certain non-zero probability realizations of ϵ

$$\int \mathbb{1}_{F[v(x),a_i,\epsilon]>c>F[v(y),a_i,\epsilon]} D[F[v(x), a_i, \epsilon]] - D[F[v(y), a_i, \epsilon]] dF_\epsilon(\epsilon) > 0, \tag{2.63}$$

which since otherwise the D is non-decreasing makes the whole integral (the $u(v(x), a_i) - u(v(y), a_i)$) strictly positive.

Then we get

$$\begin{aligned} &v(x_{i1}) - v(x_{i0}) > 0 \Rightarrow \\ &\Rightarrow E[Y_{i1} - Y_{i0}|X_i = x_i, A_i = a_i] = u[v(x_{i1}), a_i] - u[v(x_{i0}), a_i] > 0 \forall a_i, \\ &\Rightarrow E[Y_{i1} - Y_{i0}|X_i = x_i] > 0. \end{aligned} \tag{2.64}$$

[15] Basically, what we do is replace F with F_i, but since we only compare across time, the heterogeneity of F_i through the cross-section does not matter if we only focus on orders. The crucial time-homogeneity assumptions are that the functions v, u and the scalar a_i are constant. This is Manski's and Hoderlein and White's insight.

Again, making the symmetric argument starting with $v(x_{i1}) - v(x_{i0}) < 0$ and considering the trivial case with equality yields the result.

2.7.2 Proof of Lemma 10 and Proposition 4

If we have $\tilde{c}_n$ and c_n sequences that give the same level curves, it must be true that $\lambda\{\tilde{c}_n\} + (1-\lambda)\{c_n\}$ also satisfies all our restrictions from above for any $\lambda \in [0,1]$. This means that there are uncountably many sequences in any ϵ-ball around the true c_n sequence that rationalize the data. That means that we have an ill-posed problem. Using Newey (1997)'s results, we would conclude that the $\hat{\beta}$ sequence converges in probability to a continuum of sequences, which is impossible. On the other hand, if the uniqueness condition is satisfied, the arguments in Newey (1997) prove consistency.

As for the second lemma, the quasi-linear case and the additive case are well-known to be identified from micro theory. In the first case we know $\frac{\partial^2 v(x)}{\partial x^1 \partial x^1} = 0$, which is enough restriction to inductively calculate $\partial^\alpha v(x)/\partial^{(1,0,\ldots,0)} v(x)$ from the ratio of first partial derivatives, which are identified as the slopes of the level curves.

For the additive case we consider that any strictly increasing function that is additive in its argument must be an affine function. So the possible monotone transformations for the true v that preserve the level curves are all affine.

2.7.3 Proof of Lemma 12-13

For the first statement, consider that if $\Delta^m_{ijji} > 0$ for some x_i, x_j, then due to the continuity of f_m there is an ϵ-ball around (x_i, x_j) that is non-zero as well,

which then contradicts the assumption that $\Delta^m_{ijji} = 0$ a.e.

As for the second lemma, by the zero property, for any q

$$w(x + dx^q, x + dx^q) = 0 = w(x, x) + \left.\frac{\partial w(x,y)}{\partial x^q}\right|_{x=y} dx^q + \left.\frac{\partial w(x,y)}{\partial y^q}\right|_{x=y} dx^q, \tag{2.65}$$

which gives by symmetry that

$$0 = \left[\left.\frac{\partial w(x,y)}{\partial x^q}\right|_{x=y} + \left.\frac{\partial w(x,y)}{\partial y^q}\right|_{x=y}\right] dx^q = 2\left.\frac{\partial w(x,y)}{\partial x^q}\right|_{x=y} dx^q, \tag{2.66}$$

which gives the required conclusion.

2.7.4 Proof of Proposition 6

During this proof we will greatly rely on Belloni et al. (2015) and Newey (1997). We prove the consistency and the L^2 rate of convergence for the first-order variant of $\hat{\beta}$, denoted as $\tilde{\beta}$. To calculate this estimator, while maintaining the i.i.d. assumption of the (X_i, A_i) vectors, we only use one node's vector of characteristics once in the estimation process. That is, denote $\Delta_{i_4} = \Delta_{4i-3,4i-2,4i-1,4i}$ with its elementwise analogues, and $\tilde{w}_{i_4} = \hat{S}_{4i-3,4i-2}(X_{4i-1}) \cdot \hat{S}_{4i-3,4i-2}(X_{4i})$, then

$$\tilde{\beta} = \min_b \frac{N}{4} \sum_{i=1}^{N/4} w_{i_4} \left[\Delta_{i_4} b - \Delta^q_{i_4}\right]^2. \tag{2.67}$$

If we stack the $N/4$ vectors Δ_{i_4} and scalars $\Delta^q_{i_4}$, we get

$$\tilde{\beta} = [\Delta_4' \tilde{W} \Delta_4]^{-1} \Delta_4' \tilde{W} \Delta_4^q. \tag{2.68}$$

1. decomposition of $\hat{w}$

Given the sufficient variation condition (Assumption 15), we can normalize the

$$Q_0 = E\left[[\Delta_{ijji}^{q+1} \ ... \ \Delta_{ijji}^{q+M}]'[\Delta_{ijji}^{q} \ \Delta_{ijji}^{q+1} \ ... \ \Delta_{ijji}^{q+M}]\right] = I,$$

the $M \times M$ identity matrix.[16] Our target is

$$\begin{aligned}\left|\left|w - [\Delta_{ijji}^{q} \ \Delta_{ijji}^{q+1} \ ... \ \Delta_{ijji}^{q+M}] \begin{bmatrix} 1 \\ \tilde{\beta} \end{bmatrix}\right|\right| \leq & \left|\left|w[\Delta_{ijji}^{q} \ \Delta_{ijji}^{q+1} \ ... \ \Delta_{ijji}^{q+M}] \begin{bmatrix} 1 \\ \beta \end{bmatrix}\right|\right| + \\ & + \left|\left|[\Delta_{ijji}^{q+1} \ ... \ \Delta_{ijji}^{q+M}](\beta - \tilde{\beta})\right|\right| = \\ & = O(M^{-r}) + O_p(||\tilde{\beta} - \beta||),\end{aligned} \tag{2.69}$$

after the necessary normalizations on the w function. Here β are the true coefficients that correspond to the function that is the closest to w in the space spanned by the $f_q, f_{q+1}, ..., f_{q+M}$, given our assumptions and normalizations. We used the triangle inequality, and the normalization of Q_0, and the representation of w from the zero property.

2. Bias on the structural level

Our first step towards $\beta - \tilde{\beta}$ is considering that there is a non-zero constant m for which

$$\begin{aligned} & m|w(x_i, x_k) + a_i - w(x_j, x_k) + a_j| < \\ & < |F_\epsilon[w(x_i, x_k) + a_i + a_k] - F_\epsilon[w(x_j, x_k) + a_j + a_k]| \Rightarrow \\ & \Rightarrow m|w(x_i, x_k) + a_i - w(x_j, x_k) - a_j| < |S_{ij}(x_k)|. \end{aligned} \tag{2.70}$$

[16] See additional justification for this in Newey (1997).

Here we rely on the strict monotonicity of F_ϵ, and the implication that the sign of the difference on both sides in the above inequalities is the same. Denote

$$\Delta^w_{ijkl} = w(x_i, x_k) - w(x_j, x_k) - w(x_i, x_l) + w(x_j, x_l).$$

We can conclude

$$m|\Delta_{ijkl}| < |S_{ij}(x_k)| + |S_{ij}(x_l)| \tag{2.71}$$

by the triangle inequality. This gives that if $\hat{S}_{ij}(x_k), \hat{S}_{ij}(x_l) \leq CN^{-h}$, then

$$|\Delta^w_{ijkl}| < m^{-1}|S_{ij}(x_k) + \hat{S}_{ij}(x_k) - \hat{S}_{ij}(x_k)| + m^{-1}|S_{ij}(x_l) - \hat{S}_{ij}(x_l) + \hat{S}_{ij}(x_l)| \leq \tag{2.72}$$

$$\leq m^{-1}|\hat{S}_{ij}(x_k)| + m^{-1}|\hat{S}_{ij}(x_l)| + 2m^{-1}\sup_x |S_{ij}(x) - \hat{S}_{ij}(x)| =$$

$$= O\left(N^{-h}\right) + O_p\left[\left(\frac{\log N}{N}\right)^{\frac{2}{p+4}}\right] \tag{2.73}$$

under our assumption on the rate of the bandwidth σ_N according to Stone (1982).[17] From here it follows that after applying the linear approximation and plugging in $\beta^m = -\frac{c_m}{c_q}$

$$\Delta^q_{ijkl} = \Delta_{ijkl}\beta + O\left(N^{-h}\right) + O_p\left[\left(\frac{\log N}{N}\right)^{\frac{2}{p+4}}\right] + O\left(M^{-r}\right). \tag{2.74}$$

Here the third error term is the bias from the non-parametric approximation, the second one is the error from the first stage, and the first one is the error from conditioning. If $h < \frac{2}{p+4}$, we control the only unobservable random term. This

[17] This is the optimal rate.

means that we only care about the bias from now on. All the randomness due to unobservables is taken out after the first stage this way.

3. ξ_m

Before we would start to derive the result analogous to Theorem 4.1 in Belloni et al. (2015), we need to calculate $\sup ||\Delta^M_{ijkl}|| = \xi^{\Delta}_M$ (see Newey (1997) for more details). Since

$$\begin{aligned}\sup ||\Delta_{ijkl}|| &= \sup ||f(x_i, x_k) - f(x_j, x_k) - f(x_i, x_l) + f(x_j, x_l)|| \leq \\ &\leq \sup ||f(x_i, x_k)|| + \sup ||f(x_j, x_k)|| + \sup ||f(x_i, x_l)|| + \sup ||f(x_j, x_l)|| \\ &= 4 \sup ||f(x_i, x_k)|| = 4\xi_M,\end{aligned} \tag{2.75}$$

so that as m grows, $O(\xi_M) \geq O(\xi^{\Delta}_M)$.

4. Appropriate scaling for matrix LLN

Our first step is to ensure that the weighted Gram matrix is converging after some scaling. By Assumption 15 we normalize the matrix

$$Q_M = E[{\Delta^M_{ijkl}}' \Delta^M_{ijkl} | S_{ij}(X_k) = S_{ij}(X_l) = 0], \tag{2.76}$$

to the identity for every N, as they have only positive eigenvalues.[18] Then

$$||[(Nh_N^2)^{-1} \Delta_4' \tilde{W}_4 \Delta_4 - I]||_2 = o_p(1). \tag{2.77}$$

First, the quadratic form is linear in the coefficients,

$$X'(A + B)X = (X'A + X'B)X = X'AX + X'BX,$$

[18]Note that this means we orthogonalize the Δ^q_{ijkl} functions at every step, so we would need to index the f_m functions by M as in Newey; however, for the sake of brevity, we do not include this subscript.

so it is enough to prove

$$||[(Nh_N)^{-1}\Delta_4' W_4 \Delta_4 - I]||_2 = o_p(1), \tag{2.78}$$

after adding and subtracting the quadratic form using the true screening values and using the triangle inequality. As a reminder, W here is a diagonal matrix of size $N/4$ (rounded down to an integer), with $\mathbb{1}[|S_{i,i-1}(X_{i-3})| < h_N] \cdot \mathbb{1}[|S_{i,i-1}(X_{i-4})| < h_N]$.

This makes $\Delta_4' W_4 \Delta_4$ a sum of independent matrices for any M, and *eventually* as $N \to \infty$, the number of matrices in this sum is going to be equal to the effective sample size (the probability that $\mathbb{1}[|S_{i,i-1}(X_{i-3})| < h_N] \cdot \mathbb{1}[|S_{i,i-1}(X_{i-4})| < h_N] = 1$ multiplied by $N/4$). Now for some constant M

$$\begin{aligned} |S_{ij}(x_k)| &\leq E_{A_k}|F_\epsilon[w(x_i, x_k) + a_i + A_k] - F_\epsilon[w(x_j, x_k) + a_j + A_k]| < \\ &< M|w(x_i, x_k) + a_i - w(x_j, x_k) - a_j| \end{aligned} \tag{2.79}$$

by Jensen's inequality and our smoothness condition, so

$$\begin{aligned} P[|S_{ij}(X_k)| < h_N] &\geq P[|w(X_i, X_k) + A_i - w(X_j, X_k) - A_j| < M^{-1}h_N] \\ &\geq P[|w(X_i, X_k) - w(X_j, X_k)| + |A_i - A_j| < M^{-1}h_N] \\ &\geq P[|w(X_i, X_k) - w(X_j, X_k)| < M^{-1}h_N] + P[|A_i - A_j| < M^{-1}h_N] \\ &= O(h_N) + O(h_N) = O(h_N). \end{aligned} \tag{2.80}$$

This is true, since by assumption for every X_i, X_j there are some equidistant points X_{ij}^*, and $w(x_i, x)$ is uniformly Lipschitz, so for any x_i and x_j realization there is a universal constant M_2 for which

$$|w(X_i, X_{ij}^*) - w(X_j, X_{ij}^*)| - |w(X_i, X_k) - w(X_j, X_k)| < M_2||X_k^* - X_k||.$$

Alternatively, if we can assume that $Supp(w(X_i, X_k)) \subset Supp(A_i)$, then we can jump to use the strong density assumption for the joint distribution of (X_i, A_i) to come to the same conclusion.

Either way, this shows that the effective sample size, and correct scaling factor is Nh_N^2. Invoking Lemma 6.2 from Belloni et al. (2015) concludes with the desired result with the sufficient condition

$$\xi_M^2 \frac{\log M}{Nh_N^2} \to 0 \tag{2.81}$$

5. Aggregated bias

Now we proceed as usual to calculate the bias

$$\begin{aligned} \tilde{\beta} &= [\Delta_4' \tilde{W} \Delta_4]^{-1} \Delta_4' \tilde{W} \Delta_4^q \\ &= \beta + [\Delta_4' \tilde{W} \Delta_4]^{-1} \Delta_4' \tilde{W} \upsilon_4 \end{aligned} \tag{2.82}$$

where υ_4 is a $N/4$-vector of bias terms:

$$\upsilon_{4,i} = O\left(N^{-h}\right) + O_p\left[\left(\frac{\log N}{N}\right)^{\frac{2}{p+4}}\right] + O\left(M^{-r}\right).$$

We know that under the condition in equation (2.81), the Gram-matrix converges to the identity after dividing by Nh_N^2, so we are left with

$$(Nh_N^2)^{-1} \Delta_4' \tilde{W} \upsilon_4 = O\left(N^{-h}\right) + O_p\left[\left(\frac{\log N}{N}\right)^{\frac{2}{p+4}}\right] + O\left(M^{-r}\right). \tag{2.83}$$

6. Binding constraint and result

Denote $M = C_M N^m$. We need to choose h and m such that the rate of the bias terms is minimized. Let $O(\xi_M) = N^{mk}$, where k is typically 1 (e.g. polynomial

series) or 1/2 (e.g. B-splines). For now we ignore the term coming from the first stage. Then the problem becomes

$$\begin{aligned} \max_{m,h} \min(h, rm) \\ \text{s.t. } 2km - 1 + 2h < 0, \\ m, h > 0. \end{aligned} \tag{2.84}$$

For simplicity, we leave some slack for the constraint by ignoring the logarithmic term, as it is increasing slower than any power rate. Assume there is a solution (m^*, h^*) and so the constraint is not binding. Then we would be able to increase both m^* and h^* by a small enough ϵ without violating the constraint, which then contradicts that they were optimal at the first place. This also means that for the supremum value v_s and the corresponding h_s, m_s we have

$$km_s + h_s = \frac{1}{2}, \tag{2.85}$$

$$v_s = \max_m \min(0.5 - km, rm), \tag{2.86}$$

so $m_s = \frac{1}{2(k+r)}$ and $v_s = h_s = \frac{r}{2(k+r)}$. This means we want that for a small $\epsilon > 0$

$$m = \frac{1}{2(k+r)} - \epsilon r^{-1} \tag{2.87}$$

$$h = \frac{r}{2(k+r)} \tag{2.88}$$

and the achieved convergence rate in the L_2 norm can be up to $N^{-\frac{r}{2(k+r)}}$.

Note that this is generally higher than $\frac{2}{p+4}$, because typically $k \geq 1/2$ and $r \leq 1/p$, which gives

$$\frac{r}{2(k+r)} = \left[2\left(\frac{k}{r} + 2\right)\right]^{-1} \leq \frac{1}{p+4},$$

which is exactly half of the standard optimal non-parametric rate from Stone (1982). Indeed, the bias from the first stage is not binding.

7. U-statistics considerations

First, the $\hat{\beta}$ incorporates more information about the model than $\tilde{\beta}$, so we have that our original estimator will behave at least as well as $\tilde{\beta}$.

However, as we know from Serfling (1980), for example, a non-degenerate U-statistic version of a statistic is not going to converge at a faster rate. For this reason, the elementwise rate of convergence in the Gram-matrix is not going to differ. We also do not expect the rate at which the size of the matrix influences the law of large numbers to change, given the non-degeneracy. The reason for this is that the relationship between the variance of a degenerate kth order U-statistics and its first-order counterpart is a multiplying scalar that depends on the k, not the bound.

Formally, in Belloni et al. (2015), the critical step is to give a condition for which the Gram-matrix converges. Lemma 6.2 is the key, which is Rudelson's LLN for matrices. Its proof consists of a symmetrization lemma and a Khinchin inequality, and some additional algebra and estimation of the second moment by boundedness. Out of these three steps, only the symmetrization lemma requires independence, which is violated by the entries of Δ_{ijkl} if we use the forth order statistic. Note that except for this key step, the rest of the proof is the same, as we do not use any assumption about the correlation of the matrices of RHS or LHS variables.

First, note that our statistic is not a forth order U-statistic, but two second-order nested. This allows us to specialize in second-order statistics equentially. One can adopt an analogue approach to Rudelson (1999), because after a Hoeffding-decomposition we have that the first non-degenerate terms can be handled by the same Lemma 6.2 from Belloni et al. (2015), and we are only left with a degenerate case. For this the symmetrization argument from Nolan and Pollard (1987) after conditioning on the Δ^M_{ijkl}-s we can get that whatever the rate was at the empirical processes, it will be doubled in case of the second order degenerate term.

Instead of this, we will argue that the difference between the $\hat{\beta}$ and $\tilde{\beta}$ estimates is negligible compared to this slow, nonparametric rate. The scaled difference between the two objective functions

$$R_N = \binom{N}{2}^{-1} \binom{N-2}{2}^{-1} \sum_{i<j} \sum_{k<l} \hat{w}_{ijkl} \left[\Delta^q_{ijkl} - \Delta_{ijkl} b_M\right]^2 - \\ - \frac{N}{4} \sum_{i=1}^{N/4} w_{i_4} \left[\Delta_{i_4} b - \Delta^q_{i_4}\right]^2 = O_p(h_N^2) \tag{2.89}$$

under our assumptions (notably, the bandwidth). This is because the expectations cancel each other out, and the degenerate parts of the Hoeffding-decomposition will behave regularly after a scaling with ξ_M^{-2}. The kernel of the U-statistic is Euclidean as established in Sherman (1994), for example, which means that the degenerate part is $O_p(\xi_M^2 N^{-1})$ according to Nolan and Pollard (1987), Theorem 9.[19] Then we are only left with the empirical processes, which are $O_p(\max(\sqrt{Nh_N^2}, Nh_N^2 \xi_M \log(M)))$, exactly along the argument presented in the previous part. If we choose the rate

[19] We already undid the scaling.

for M, h_N as prescribed above, we get $O(h_N^2) \geq O(\sqrt{Nh_N^2}^{-1})$, hence the result. The objective function for $\hat{\beta}$ is not a second order process, it is more like a 2×2 U-process, so one needs to repeat this argument once more. However, the rates are only going to get better.

From here we recast some of the arguments in Pakes and Pollard (1989) in some sense, in this special case. We have two M-estimators, with objective functions that are converging faster than the convergence rate of the consistent first estimator. This implies that the second estimator is consistent too. Moreover, the rate of convergence of the second estimator will inherit the first rate.

In our case, if we choose m and h as prescribed above to satisfy the criterion for $\tilde{\beta}$ to be consistent, we have that $O_p(R_N) = \frac{-r}{k+r} + \epsilon$. This is a faster rate than $\frac{-r}{2(k+r)} + \epsilon$ under our condition in the proposition, so the $\hat{\beta}$ is consistent. Also,

$$\begin{aligned} ||\beta - \hat{\beta}|| \leq ||\beta - \tilde{\beta}|| + ||\hat{\beta} - \tilde{\beta}|| = \\ = O_p\left[N^{-\frac{r}{2(r+k)}+\epsilon}\right] + O_p\left[N^{-\frac{2k+r}{4(r+k)}+\epsilon}\right] = O_p\left[N^{-\frac{r}{2(r+k)}+\epsilon}\right]. \end{aligned} \tag{2.90}$$

To see this last step, for the ease of notation, we will define the two functions

$$f(X, b) = (Xb - y)'(Xb - y), \tag{2.91}$$

$$g(X, b) = f(x, b) + O_p(R_N), \tag{2.92}$$

for $x^i \in \mathbb{R}^s$, $y \in \mathbb{R}$, $b \in \mathbb{R}^s$ and the two estimator

$$\hat{b} = \min_b f(X, b) \tag{2.93}$$

$$\tilde{b} = \min_b g(X, b) \tag{2.94}$$

to create full analogy with the estimators above. We assume both f and g have a unique minimum. Let us have that

$$f(X, \hat{b}) = R_{2N}$$

is the same as $||\hat{b} - b_0||^2$, the true vector.

Then by the definition of $\hat{b}$ and $\tilde{b}$, we get

$$0 < f(X, \tilde{b}) - f(X, \hat{b}) < O_p(R_N). \tag{2.95}$$

This means that

$$0 < (X(\hat{b} + \tilde{b}) - 2y)' X(\tilde{b} - \hat{b}) < O_p(R_N). \tag{2.96}$$

Now we need a reverse Cauchy-Schwarz inequality result; we choose the one from Dragomir (2015),[20] which results in

$$\begin{aligned} 0 \leq ||2y - X(\tilde{b} + \hat{b})||^2 ||X(\tilde{b} - \hat{b})||^2 - |(X(\hat{b} + \tilde{b}) - 2y)' X(\tilde{b} - \hat{b})|^2 \leq \\ \leq 4||y - X\hat{b}||^2 ||X(\tilde{b} - \hat{b})||^2, \end{aligned} \tag{2.97}$$

if we choose $a + A = -2$ and $\frac{|A-a|}{2} = \frac{2||y - X\hat{b}||}{||X(\tilde{b} - \hat{b})||}$ in his Theorem 1.[21] After rearranging we have

$$0 \leq \left[||y - X(\tilde{b} + \hat{b})/2||^2 - ||y - X\hat{b}||^2 \right] ||X(\tilde{b} - \hat{b})||^2 \leq O_p(R_N^2), \tag{2.98}$$

[20] There must be a more elementary way to prove the following line.

[21] This selection is always feasible as long as $\tilde{b} \neq \hat{b}$ and the matrix X is full-rank s - these two are considered to be nuisance cases.

where we know that the expression in the middle is positive $\hat{b}$ is minimizing f (and not $(\hat{b}+\tilde{b})/2$). Note that this cannot happen, unless $||\hat{b}-\tilde{b}|| \to 0$. Then since $y - X(\tilde{b}+\hat{b})/2||^2 = ||y - X\hat{b} - X(\tilde{b}-\hat{b})/2$, the term in the middle will either have the rate $||X(\hat{b}-\tilde{b})||^4$ or we have that $||X(\hat{b}-\tilde{b})||^2 \geq O(R_{2N})$, the doubled rate of convergence for $\hat{b}$. Choose $R_{3N} = \max(R_{2N}, R_N)$, then

$$(\tilde{b}-\hat{b})'X'X(\tilde{b}-\hat{b}) \leq O_p(R_{3N}). \tag{2.99}$$

Now we know we need to scale $X'X$ with some T_N sequence in order it to converge to the identity. Then

$$||\hat{b}-\tilde{b}|| = O_p\left(\sqrt{R_{3N}T_N^{-1}}\right). \tag{2.100}$$

In our case $T_N = Nh_N^2 = N^{\frac{k}{k+r}}$, while the rate for the bandwidth has to satisfy the conditions mentioned above. Also, the rate of R_{3N} is $-\frac{r}{2(k+r)}$, as h_N^2 is twice as high as the convergence rate, which gives

$$||\hat{\beta}-\tilde{\beta}|| = O_p\left(N^{-\frac{r+2k}{4(k+r)}}\right).$$

This gives us that if the first-order statistic converges, the estimator based on the U-statistic is not going to be worse. Using the triangle inequality in the other direction also lets us prove that the estimator based on the U-statistic will not attain the optimal non-parametric rate. Assume it does, then the estimator based on the empirical process would need to converge at a higher rate than we saw possible. (We need for this that the LLN for the matrix gives necessary conditions for the convergence of the Gram-matrix.)

Chapter 3

The empirical content of the Nash-equilibrium assumption in discrete games

3.1 Introduction

In this paper we examine the empirical content of the assumption that in a complete information game the players play pure strategy Nash-equilibrium. In particular, we focus on a two-player game with payoff-functions

$$\pi_i(s_i, s_{-i}) = \begin{cases} \alpha_i - \Delta_i s_{-i} + \beta_i x_i + \epsilon_i & \text{if } s_i = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

where the 2-vector $s = (s_1, s_2)$ are the observed outcomes, the possible actions of the game $s_i \in \{0, 1\}$, the $x = [x_1\ x_2]$ matrix with size $m_1 + m_2 \times 1$ is the observable regressors, and $\epsilon = (\epsilon_1, \epsilon_2)$ is the vector of unobservables. We assume that the econometrician does not observe the ϵ vector, but the two players have complete information. We refer to the β_i as the slope parameters, and the α_i as the intercept. The main topic of this paper is strategic interaction effect, the vector $\Delta = (\Delta_1, \Delta_2)$, which is the most difficult parameter to identify in this model. Δ is the change in the payoff when the other Player starts to play 1 instead of 0. If the game above is an entry game between to firms, this is the loss of profit firm i suffers because it has to share the market with the other firm. If Δ is small, it means that strategic considerations do not play a major role in the firm's decision to enter the market

or not. A crucial assumption of this paper is that the strategic interaction effects have identical signs (positive) for both players.

The game above is important for many applications, most notably in the field of industrial organization literature while modeling the firm's entry decision. If the parameter of strategic interaction (Δ) is positive, then the game describes the situation when two firms make simultaneous decisions whether to enter a given market or not. The entry game is of great importance in the field of industrial organization, and the estimation of the interaction parameters has a long literature in econometrics. In the followings I will remain with this important example, even though the same type of argument would work for the case when the interaction effects are all negative.[1] The pure strategy assumption[2] is consistent with the real life cases when there is no coordination failure in the sense that the firms would not play $(0, 1)$ or $(1, 0)$ because they would be Pareto-dominated. Besides modeling the entry decision, the form of the game above appear among others in models of political competition, labor force participation (Soetevent and Kooreman (2007)) or product differentiation (Mazzeo (2002)).

The contribution of this paper is the identification of the interaction parameters up to scale without large support assumption under weak and testable conditions on the shape of the distribution of the unobservables. While presenting the identification argument, we show that the Nash assumption restricts the joint cumulative distribution (cdf) and the joint probability density (pdf) of the unob-

[1]The differences between the two cases are described in Tamer (2003).
[2]following Kline (2015) and Bresnahan and Reiss (1991)

servables in a way that the identification problem can be described by a simple version of an image stitching problem (see for example Szeliski 2006). To achieve point identification, we make the usual econometric assumptions of the presence of continuous observables and strong exogeneity. As an additional key condition, we assume that some upper- or lower-contour sets of the joint pdf of the unobservables are strictly convex on the identified region. This set of assumptions is sufficient to achieve point identification, but does not rely on identification-at-infinity type arguments. Moreover, the restriction on the shape of the distribution of the unobservables includes classes of distributions that are ruled out by previous results such as Kline (2015). The condition on the shape of the distribution is testable in the sense that we only make assumptions on identified objects.

The issue of identification in complete information games has been addressed by many papers before, good survey articles are for example Berry and Tamer (2006) or Berry and Reiss (2007). The problem was formulated by Bresnahan and Reiss (1991a,b) as an entry game between two firms. In this game the firms decide if they enter a market or not. If a firm enters alone, it will get a higher profit compared to the case when both firms enter, just like in the payoff functions above. The main challenge in this literature is to handle the incompleteness of the (economic) model: the region of multiple equilibria. For example in the entry game defined in the previous section, for some payoff parameters the $(0, 1)$ outcome is a Nash-equilibrium just as the $(1, 0)$ outcome.

The literature had three ways to circumvent the identification problem. Examining the entry game, Bresnahan and Reiss (1991a) observe two facts. First,

while we do not always know which firm enters the market, we know how many of them do so. Second, for some parameter values, there is only one equilibrium. For example when the profits are negative or positive regardless if the firm enters alone or with a competitor, we know that both firms will stay away or enter, respectively. Based on this, the authors make use of "unique potential outcomes", which are the $(1, 1)$ and $(0, 0)$ outcomes in the entry game above. The "unique potential outcomes" are the pair of actions that occur if and only if the value of the unobservables are realized in a given subset of their joint support. Given only this information, the identification of the slope parameters is possible. As Kline (2015a) shows, after putting shape and location restrictions on the distribution of the unobservables, (two) unique outcomes give identification of the interaction parameters, even if the observables are of bounded support.

Second, Bajari et al. (2010) solve the problem of multiple equilibria by assuming a randomization device. If the payoffs are realized in the range of multiple equilibria, this randomization device would every time decide according to a parametrized random variable which equilibrium the game lands at. This solution comes at a cost of further exclusion restrictions regarding the device, but can also handle mixed strategies or the cases without unique outcomes, unlike fir example this present paper. This line of literature can be characterized as putting restrictions on equilibrium selection, and is very popular in the IO literature (for example Fox and Lazzati 2015).

Third, Tamer (2003) uses an approach that requires a regressor with a large support to identify the slope and interaction parameters up to scale. His approach is

suitable for the case without unique outcomes and for handling mixed actions. Here the effect of the regressor needs to dominate the unobservables, which is satisfied at infinity. This may not be a very plausible assumption, and the (theoretical) demand for the relaxation of the large support conditions has appeared already in Ciliberto and Tamer (2009). In that paper the authors exploited that the probabilities of the different equilibrium outcomes must add up to one, and achieved estimation by partial identification. Kline (2015b) utilizes the large support assumption to achieve semiparametric identification of the payoff functions even after weakening the exogeneity assumption.

This paper builds on the framework developed by Ciliberto and Tamer (2009) and Kline (2015a), as these two papers were the only ones in the literature (up to my knowledge) that considered identification without the a large support assumption (identification at infinity).

In the next section we restate the identification of Δ as a photo stitching problem. Further, we examine a sufficient condition on the joint pdf of the unobservables that provides point identification of the interaction effects. The remaining of the paper discusses simple ways how the sufficient assumption may fail and satisfied, compares the result to Kline (2015), and briefly addresses the question of the identification of the intercept. For the sake of completeness, in the Appendix we show an identification argument for the slope parameters, and define the estimator corresponding to our identification argument.

Payoff of Player 1	Payoff of Player 2	Equilibrium outcome (s_1, s_2)
$\alpha_1 + \beta_1 x_1 + \epsilon_1 < 0$	$\alpha_2 + \beta_2 x_2 + \epsilon_2 < 0$	$(0,0)$
$\alpha_1 + \beta_1 x_1 + \epsilon_1 < 0$	$\alpha_2 + \beta_2 x_2 + \epsilon_2 \geq 0$	$(0,1)$
$\alpha_1 + \beta_1 x_1 + \epsilon_1 \geq 0$	$\alpha_2 + \beta_2 x_2 + \epsilon_2 < 0$	$(1,0)$
$\alpha_1 + \beta_1 x_1 - \Delta_1 + \epsilon_1 < 0$ $\alpha_1 + \beta_1 x_1 + \epsilon_1 \geq 0$	$\alpha_2 + \beta_2 x_2 - \Delta_2 + \epsilon_2 < 0$ $\alpha_2 + \beta_2 x_2 + \epsilon_2 \geq 0$	$(0,1)$ or $(1,0)$
$\alpha_1 + \beta_1 x_1 - \Delta_1 + \epsilon_1 \geq 0$	$\alpha_2 + \beta_2 x_2 - \Delta_2 + \epsilon_2 \geq 0$	$(1,1)$

Table 3.1: Pure strategy Nash-equilibria of the entry game (the interaction effects are assumed to be positive).

3.2 Econometric model

3.2.1 Predictions of the economic model

Abusing the notation above sightly, the economic model yields the following the best responses for $i \in \{1, 2\}$

$$s_i = \begin{cases} 1 & \text{if } \pi_i(1, s_{-i}) \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

The predictions of the economic model can be summarized by the set of Nash-equilibria, which depends on the realization of the ϵ vector for the given observables, X. Table 3.1 describes the possible equilibrium outcomes. There is a region of multiple equilibria, which creates the identification problem. However, after assuming that the players play Nash-equilibria, the $(0,0)$ and $(1,1)$ outcomes happen if and only if the unobservables fall into the extreme low or high regions described by the first and last rows of Table 1. The graphical version of this table is Figure 2 from Tamer (2003).

3.2.2 Econometric assumptions

The data that the researcher observes consists of the observables (X) and the outcome of the game (s), as described by (3.2). I assume that the same game is being played multiple (infinitely many) times, with the unobservables and observables always redrawn from the same joint distribution. In the leading example this usually means that the researcher observes many different independent markets with the two firms.

The slope parameters are usually identified to scale only, for example with the normalization that the first entry of the vector is one ($\beta_i[1] = 1$) (see for example Kline 2015b). This also means that the identification results in this paper are also up to the same scale. This scaling should not interfere with our argument, assuming that it did not change the sign of the interaction effects. This complication is less severe because the argument in de Paula and Tang (2012) for the identification of the signs of the parameters is applicable for this case as well. Now we state assumptions under which the slope parameters are identified according to Kline (2015).

Assumption 17 ($\beta_i^1 = 1$)

The first entry of the vector of slope parameters for every player is 1.

The usual key econometric assumptions are continuity, exogeneity and a sufficient variation assumption on X.

Assumption 18 (Continuous X)

For every player theres is at least one observable with a non-zero slope coefficient that is a continuous random variable.

Assumption 19 (Exogeneity)

X and ϵ are independent of each other. Specifically, conditional on any values of X, the vector of unobservables have the same joint distribution.

Assumption 20 (Sufficient variation of X)

The support of X is not a proper subspace of $\mathbb{R}^d$.

We use a shorthand for the probability that the first player plays a and the second b: $P[s_1 = a, s_2 = b]$ is denoted as $P[(a, b)]$. Moreover, given the special role the intercept parameters α_i have, it is worth abstracting from them, and so I define

$$\tilde{\epsilon}_i = \epsilon_i + \alpha_i.$$

As a technical assumption, I also assume that the $\tilde{\epsilon}$ has a non-degenerate *continuous* density.[3]

Assumption 21 (ϵ has a continuous non-degenerate density.)

A continuous version of the probability density function of $\tilde{\epsilon}$ exists and it is denoted as $f_{\tilde{\epsilon}}$.

After the β-s are already identified, we can abstract from them as well, and

[3]Such a density is defined as a Radon-Nikodym derivative with respect to the Lebesgue-measure on $\mathbb{R}^2$. The argument works without continuity as well, but the explanation of the results are much simpler.

focus on the variation of the observable part of the score. I define

$$c_i(x) = -\beta_i x_i,$$

the observed part of the index for player i, which includes all the information the firm needs from the observables to make an informed decision. Throughout the paper we use the notation $c = (c_1, c_2)$, a 2-vector.[4] We need the vector of c-s to vary across the observed games. This is incorporated in Assumption 22 below. In this paper, the support of a random variable is the set of values that the random variable would take with non-zero probability, and it is denoted as $Supp(.)$.

Assumption 22 (Support of c)

The support of the vector $c = [c_1, c_2]$ contains a rectangle on $\mathbb{R}^2$, such that it is the product of two closed intervals on $\mathbb{R}$.

The supremum of such rectangles is S, a Cartesian-product of the intervals $[\underline{c}_1, \bar{c}_1]$ and $[\underline{c}_2, \bar{c}_2]$:

$$S = [\underline{c}_1, \bar{c}_1] \times [\underline{c}_2, \bar{c}_2] \in Supp(c).$$

Further define the dimensions of this largest rectangle S as

$$\bar{c}_1 - \underline{c}_1 = h_1,$$

$$\bar{c}_2 - \underline{c}_2 = h_2.$$

Assumption 22 is relatively simple to test after the slope parameters are known, since one only needs to see if the observable indexes are on the same line (or point) or not.

[4] just as $\epsilon = (\epsilon_1, \epsilon_2)$, and so on.

We need to assume that the set of possible interaction parameters is compact. This is stated in Assumption 23 below.

Assumption 23 (The set of possible Δ.)

Assume that $\Delta \in D$, a rectangle in $\mathbb{R}^2$ with dimensions d_1 and d_2 are strictly smaller than the respective h_i. That is,

$$0 \leq \Delta \leq d = (d_1, d_2) << h = (h_1, h_2).$$

That is, I assume that the set of possible interaction effects forms the rectangle D, with dimensions d_1 and d_2. Writing this with inequalities, I assume that

$$0 \leq \Delta_i \leq d_i.$$

As mentioned above, there is an additional necessary condition needed concerning the set of possible Δ's. The problem is that the support of the c vector may be smaller than the set D. In that case, there is no observable difference between two data generating processes with strategic interaction parameters that are greater than the respective h_i. This is amended by Assumption 23.

3.3 Identification argument

In this section we identify the interaction parameters.

3.3.1 The joint probability distribution function and the Nash assumption

First, I argue that the conditional probabilities of the unique outcomes are point-identified by the model. From Table 3.1, these probabilities are

$$\begin{aligned} P[(0,0)|X] &= P[\alpha_1 + \beta_1 x_1 + \epsilon_1 < 0, \alpha_2 + \beta_2 x_2 + \epsilon_2 < 0|X] = \qquad (3.3) \\ &= P[\tilde{\epsilon}_1 < c_1, \epsilon_2 < c_2 | c_1 = -\beta_1 x_1, c_2 = -\beta_2 x_2, X] = F_{\tilde{\epsilon}|X}(c_1, c_2) = \\ &= F_{\tilde{\epsilon}}(c_1, c_2) \\ P[(1,1)|X] &= P[\alpha_1 + \beta_1 x_1 - \Delta_1 + \epsilon_1 \geq 0, \alpha_2 + \beta_2 x_2 - \Delta_2 + \epsilon_2 \geq 0|X] = \qquad (3.4) \\ &= P[\tilde{\epsilon}_1 \geq c_1 + \Delta_1, \tilde{\epsilon}_2 \geq c_2 + \Delta_2 | c_1 = -\beta_1 x_1, c_2 = -\beta_2 x_2, X] = \\ &= \bar{F}_{\tilde{\epsilon}|X}(c_1 + \Delta_1, c_2 + \Delta_2) = \bar{F}_{\tilde{\epsilon}}(c_1 + \Delta_1, c_2 + \Delta_2). \end{aligned}$$

by the exogeneity assumption. Here I needed that the payoff function is additively separable in terms of β, Δ and the unobservables. This assumption is critical to be able to restate the identification problem as in section 3.3.

Since we know the conditional distribution of the outcomes, equation (3.3) gives that the joint cdf of the $\tilde{\epsilon}$ is identified on S. On the other hand, equation (3.4) shows that the conditional probabilities of the $(1,1)$ outcomes directly identify the "joint survival function" ($\bar{F}_{\tilde{\epsilon}}$) on the $S'_\Delta = S + \Delta$ set.[5] Figure 3.1 gives a visualization for the important notation from the previous and present sections (for the true Δ from D).

[5] If there is a set $A \in \mathbb{R}^2$ and a 2-vector v, then $A'_v = A + v$ means that A'_v is the translation of A by the v vector. That is, $A'_v = \{a \in \mathbb{R}^2 : a - v \in A\}$.

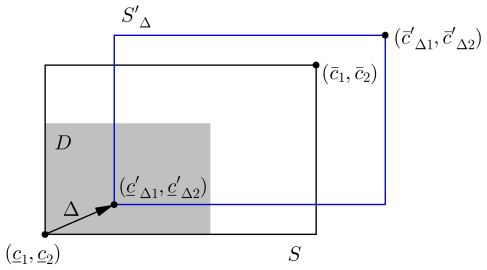


Figure 3.1: Notation summarized. The set S is the considered support of c, and the set of the possible Δ vectors is in the gray set D (starting from $\underline{c}$). On this figure, I also drew the vector corresponding the true Δ_i values (and denoted it as a Δ vector). The S'_Δ rectangle is the translation of the S by Δ. As argued, the $(0,0)$ conditional probabilities determine the cdf of $\tilde{\epsilon}$ on S, and the $(1,1)$ conditional probabilities the "survival probabilities" on the S'_Δ rectangle.

3.3.2 Identification of the strategic interaction vector

From the previous section we know that the joint cdf of the $\tilde{\epsilon}$ is identified on S by the conditional probabilities of the $(0,0)$ outcomes, and the conditional probabilities of the $(1,1)$ outcomes identify the "joint survival function" ($\bar{F}_{\tilde{\epsilon}}$) on the $S'_\Delta = S + \Delta$ set. This suggests that if we could link the joint cdf to the joint survival function, then the two equations in the previous section would tell something about the true value of the interaction effect. Unfortunately, in $\mathbb{R}^2$ the joint survival and cumulative distribution functions are only loosely connected (their sum cannot be larger than 1). However, both $F_{\tilde{\epsilon}}$ and $\bar{F}_{\tilde{\epsilon}}$ generate a joint density on the interior of their respective sets, so if they are identified on a common set, they must generate the same density there.

First, we prove that the conditional probabilities of the $(0,0)$ outcomes identify the density of $\tilde{\epsilon}$ on S.

Lemma 14

Under standing assumptions, $P[(0,0)|c]$ identifies $f_{\tilde{\epsilon}}$ on S.

Proof. Only the beginning of the proof is given here. A more technical finish is in the Appendix.

For any rectangle B on S such that

$$B = [\underline{c}_1^B, \bar{c}_1^B] \times [\underline{c}_2^B, \bar{c}_2^B],$$

that is, for $\underline{c}^B \geq \underline{c}$ and $\bar{c}^B \leq \bar{c}$ we have that

$$\begin{aligned} P[\tilde{\epsilon} \in B] =& P[(0,0)|c_1 = \bar{c}_1^B, c_2 = \bar{c}_2^B] - P[(0,0)|c_1 = \bar{c}_1^B, c_2 = \underline{c}_2^B] - \\ &- P[(0,0)|c_1 = \underline{c}_1^B, c_2 = \bar{c}_2^B] + P[(0,0)|c_1 = \underline{c}_1^B, c_2 = \underline{c}_2^B]. \end{aligned} \tag{3.5}$$

And so the probability of any such event is identified solely from knowing the probabilities of the $(0,0)$ outcomes conditional on c. Then the probability of every set that can be approximated by countably many union and intersection of these type of rectangles is also identified. That means that given the information from the data and the model assumptions, we can draw an arbitrary granular two-dimensional histogram for $\tilde{\epsilon}$ on S. On the limit, this will give the density, as it is assumed to exist. □

Given an admissible Δ, the following corollary is a direct consequence of Lemma 1.

Corollary 2

Under standing assumptions, $P[(1,1)|x]$ identifies $f_{\tilde{\epsilon}}$ on

$$S'_{\Delta} = [\underline{c}_1 + \Delta_1, \bar{c}_1 + \Delta_1] \times [\underline{c}_2 + \Delta_2, \bar{c}_2 + \Delta_2] =$$
$$= [\underline{c}'_{\Delta 1}, \bar{c}'_{1\Delta}] \times [\underline{c}'_{2\Delta}, \bar{c}'_{2\Delta}],$$

or following the notations from earlier, on $S'_{\Delta} = S + \Delta$.

This section showed that the conditional probabilities of the two unique potential outcomes $(0,0)$ and $(1,1)$ identify the joint probability density of $\tilde{\epsilon}$ on two corresponding congruent sets S and S'_{Δ}. On the other hand, we know that the second set (S'_{Δ}) is the translation of the first set (S) by the unknown vector Δ. These two sets have a non-zero measure intersect, because $d << h$. Taken the information identified up to this point, the researcher has two equally-sized rectangles and a surface above each of these sets. We know that one rectangle is the shifted version of the other, and if the translation is done with the true value of the interaction effect, the surfaces (densities) predicted by the two sources of conditional probabilities must align exactly above the implied intersect.

For identification of the Δ, we need to make sure that there is only one admissible vector $\delta \in D$ such that the densities over the implied $S \cap S'_{\delta}$ exactly align. If this is true, then this unique δ should be the true value (Δ).

3.3.2.1 Photo stitching and identification

The identification problem is a simple version of photo stitching (see a review from Szeliski 2006). If a microbiologist would like to take a detailed picture of an

organism, or an operator of a satellite of a greater area, they would take a picture of one part of the object, then move the microscope/satellite to another location, and take another photo with an overlapping region. To get a large picture, they would then stitch together all the pictures previously taken using the overlaps that would identify the relative locations of the mosaics.[6] Many times the exact optical flow, the path along they moved the satellite/microscope is not given, and someone (or rather something) needs to find along which sets the photos could be stitched together. A more complicated version of the problem is when a tourist would like to take a panorama picture of a view. The tourist would probably take a couple of photos with overlapping regions, and feed them to a program. The photo stitching application would recognize that (some linear transformation of) the patterns on the periphery of the photos are matching up to a negative distance, and stitch the photos together given the largest alignment. Figure 3.2 has an example of photo stitching from Brown and Lowe (2007).

What could go wrong with a stitching problem? When is the negative distance not identified? Non-identification can only occur when there is some periodicity in the photo. For example, consider the case when there are three exactly identical mountains in a panorama view. If the tourist took his photos such that the first photo has two of the mountains on its right side, and there are also two identical mountains on the left side of the second photo, then the stitching program could not decide if there were two or three identical mountains in reality.

[6]Hence the other name used in computer science for this kind of problems: mosaicking/mosaiking.

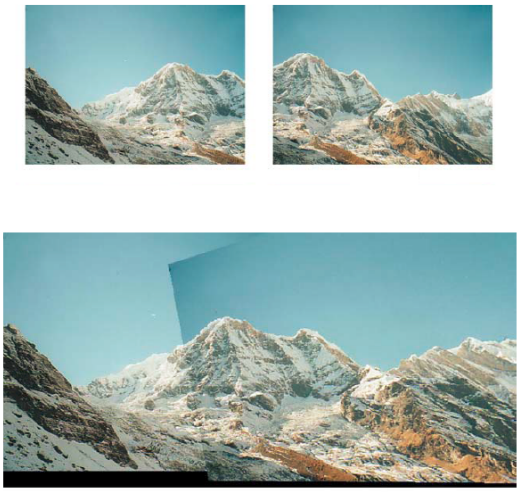


Figure 3.2: A photo stitching problem and its solution. This example is from Brown and Lowe (2007). The first two pictures are photos taken of different parts of the same view, and the lower picture is the panorama photo after the inputs are stitched together.

However, if there is a landmark on either of the mountains that breaks the perfect periodicity, we have identification once again. Another possibility is that the tourist tells the stitching algorithm that there were three mountains in the real view.

In our problem, the density is perfectly described by a two dimensional picture, for example a gray-scale representing the level (value) of the joint probability density at some point. We have two perfect "photos" of the density. One is of the set S, and one is taken of the set S'_{Δ}. The optical flow between the photos of the joint density can be described by the vector Δ, which is strictly smaller element-wise than the dimensions of S. This means that there is an overlapping region between the area of the two "takes".

Just like in any other photo stitching problem, the identifying power of the data depends on the size of the overlapping region and on the variability of the identified $\tilde{\epsilon}$ density values, which determines how many "landmarks" are on the overlapping region. We have already argued that there is a positive-sized overlap between S and S'_{Δ}, now we need to make assumptions on the *identified* regions of $f_{\tilde{\epsilon}}(t)$. Following the intuition of the photo-stitching problem, we can do two things. On the one hand, we can rule out the confusing periodicity in the density at the first place (assuming there is always at least one landmark feature), or assume how many times the same pattern from the overlapping region may occur (assuming there is 3 mountains). Even though it is harder to formulate mathematically, the former assumption is restricting an identified object, as opposed to the latter, therefore we choose to give that kind of condition. Assumption 24 states that *if* the identified density pattern in the upper-right corner of S repeats itself *twice* in

the density found above S'_Δ, *and* the density pattern in the lower-left corner of S'_Δ repeats itself *twice* on S, *then* the relative position of the repeating sets is different on S'_Δ and S. If there are no such repetitions in the density (or only on S/only on S'_Δ) at the first place, then Assumption 24 is still satisfied.

To formulate the missing assumption, it is practical to define the two "takes" of the density as separate objects, and superpose the sets S and S'_Δ. Let us define for $c \in S$

$$g^1_{\tilde{\epsilon}}(c) = -\frac{\partial^2 P[(1,1)|c]}{\partial c_1 \partial c_2},$$

which is the density identified by the $(1,1)$ conditional probabilities. We know that the above differential exists, and it is equal to $f_{\tilde{\epsilon}}(c+\Delta)$. Similarly, let

$$g^0_{\tilde{\epsilon}}(c) = -\frac{\partial^2 P[(0,0)|c]}{\partial c_1 \partial c_2},$$

the density identified by the $(0,0)$ conditional probabilities, which is the same as $f_{\tilde{\epsilon}}(c)$ from Lemma 14 again. Also, after the superposition, we call the image of the "lower-left corner" of S'_Δ, now the lower-left corner of S as G, which is defined as

$$G = [\underline{c}_1, \underline{c}_1 + s_1 - d_1] \times [\underline{c}_2, \underline{c}_2 + s_2 - d_2],$$

a rectangle with sides $s_1 - d_1$ and $s_2 - d_2$. The "upper-right corner" of S is denoted as R, another rectangle of size $(s_1 - d_1) \times (s_2 - d_2)$ defined as

$$R = [\underline{c}_1 + d_1, \bar{c}_1] \times [\underline{c}_2 + d_2, \bar{c}_2].$$

Figure 3 shows a typical G and R on S. G and R are special, because they are the largest sets in the "lower-left" and "upper-right corner" such that they are

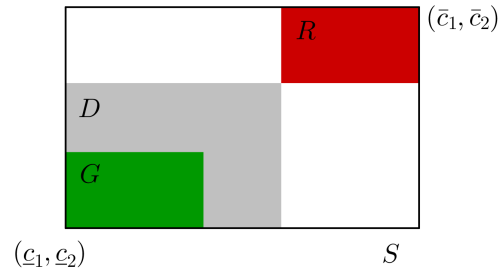


Figure 3.3: Notation summarized. The red set is R, the green is G, while the set of the possible Δ vectors is in the gray set D (starting from $\underline{c}$). The green and the red rectangles are congruent with sides $s_1 - d_1$ and $s_2 - d_2$.

subsets of all the implied intersects of S and S'_Δ, whatever value the true interaction effect takes from D.

Assumption 24 (Variation of $f_{\tilde{\epsilon}}$)

There exists no $\delta^1 \neq \delta^2$ vectors in D such that

$$\sup_{c \in G} |g^1_{\tilde{\epsilon}}(c) - g^0_{\tilde{\epsilon}}(c + \delta^1)| + |g^1_{\tilde{\epsilon}}(c) - g^0_{\tilde{\epsilon}}(c + \delta^2)| = 0 \tag{3.6}$$

$$\sup_{c \in R} |g^0_{\tilde{\epsilon}}(c) - g^1_{\tilde{\epsilon}}(c - \delta^1)| + |g^0_{\tilde{\epsilon}}(c) - g^1_{\tilde{\epsilon}}(c - \delta^2)| = 0 \tag{3.7}$$

If this assumption fails, and there were two such vectors described above, then they would give

$$R'_i = R - \delta^i, \tag{3.8}$$

$$G'_i = G + \delta^i \tag{3.9}$$

sets for $i \in \{1, 2\}$, above which the conditional probabilities of the unique outcomes would generate the same densities as they identify above R and G, respectively. Also, the (position) vector that translates G'_1 into G'_2 would be equal to the vector

that shifts R_1' into R_2', so Assumption 24 is a restriction on the degree of periodicity, indeed.

Remark.{ Assumption 24 rules out the uniformly distributed ϵ with independence as the most important special case. In fact, we do not have identification if the density is a hyperplane of $\mathbb{R}^3$ (a plane) on the sets where it is identified. Aside from the densities that coincide with hyperplanes around S, it is quite hard to come up with a realistic example that would not satisfy Assumption 24. It is relatively difficult to imagine such exact periodicity in the joint density of the unobservables on potentially sizable sets. Non-identification with a non-trivial density[7] not only requires some periodicity, but also that the set of observed indexes ended in the periodic region, and there is at least one period on both sides of the border. Holding D constant, the higher degree of periodicity is present in the ϵ's pdf, the more likely that S is such that the researcher ends up with non-identification. The densities with linear isodensity curves are the limiting case, where there is no identification for any bounded set that is smaller than the support of the unobservables. That is, with a differentiable density that is not a hyperplane of $\mathbb{R}^3$, if we encounter exact periodicity, the Nash-assumption still narrows down the possible values of the interaction vector to a finite set according to the photo-stitching intuition. }

Proposition 7 (Identification of interaction terms.)

Given Assumption 17-24, Δ is identified.

The proof of Proposition 1 is included in Appendix A, and closely follows

[7] in the sense that it is not (a) plane

the intuition outlined in this section.

An important limiting case of the identification problem is when the support of the c tends to infinity, that is, encompasses the whole plane. The following corollary highlights the relationship between the identification-at-infinity type arguments (like in Tamer 2003) and the above assumptions.

Corollary 3

For every bounded D (set of possible parameters values for Δ) under Assumption 1-6 the interaction effects are identified in the limiting case when the support of c tends to be the whole $\mathbb{R}^2$ space (infinity).[8] *That is, Assumption 24 becomes vacuously true.*

The proof of the Corollary is included in Appendix A. It is based on the observation that as the support grows, the G set becomes larger as well, because the set of possible values for the interaction effect remains the same. Then if Assumption 24 is violated, one will not be able to place two sets of size G on the S without letting them overlap each other. In fact, the area of the overlapping region must also tend to infinity. It is easy to see that the density on the intersect of the two repeating sets must be perfectly periodic. However, there exists no such density with unbounded support that is perfectly periodic on an unbounded and non-zero measure set as G. This is true because such density would never integrate

[8] Actually, it is enough if the Lebesgue-measure of S tends to infinity, but the take-away of the argument of the proof does not really change, one only needs to be cautious with directions. Also, note that this is a "pointwise" result, for every fixed D.

to 1. So the joint pdf must have either finite support, in which case identification is trivially satisfied, or Assumption 24 cannot be violated.

Remark.{ The scaled value of α_i is not identified without further assumptions. But after assuming for example

$$E[\epsilon|\epsilon \in S] = 0,$$

it is possible to give the α vector a value. Note that Assumption 24 is only restricting the shape (periodicity) of the joint density of ϵ, not the location. The location restriction is only needed for the identification of the intercept parameter.}

3.4 Simple examples for sufficient assumptions

Due to the obscure and high-level nature of Assumption 24, one may want to find more commonly used restrictions. The less useful part of the assumption is that the periodicity on S and S'_Δ must be of the same "frequency" (the repeated patterns are of the same distance from each other), because it depends on the vectors δ^2 and δ^1. So we will focus on assuming away any kind of periodicity either of the pattern found on G or R.

Going back to the photo-stitching problem, the analogue of the photo is the two-dimensional representation of the joint density. To get such a form of the pdf, I define the isodensity set (I_p) as a set of points with the same density level:

$$I_p = \{c_p \in \mathbb{R}^2 : f_{\tilde{\epsilon}}(c_p) = p\}$$

for a density level $p \in [0, 1]$. The solution of the photo-stitching problem is matching all the isodensity sets of S and S'_Δ on the true intersect. One could exploit a number

of characteristics of the isodensity sets to get assumptions that imply Assumption 24 by restricting periodicity. For example,

- how many connected components a given set consists of (except the empty set),
- (generalized) slope of the isodensity,
- the "steepness" of the pdf (the gradient).

Remark.{ In general the level sets of the pdf are identified, and there is a literature in statistics that is concerned about these kind of estimation problems (see for example Tsybakov 1997 or Nowak et al. 2009). This means that even though the assumptions will be stated on $f_{\tilde{\epsilon}}$, they are testable, since they will restrict the density only on the S and on the observed S'_{Δ}, given Assumption 24's local nature. }

3.4.1 Restrictions on peaks

In this section the restriction comes from matching peaks, or how many times the density reaches (record) low or high.

Assumption 25 (Modality of ϵ.)

There exists a level p of $f_{\tilde{\epsilon}}$ such that I_p on R (G'_{Δ}) is contained by a closed set on the interior, where it has $n > 0$ connected components. Moreover, I_p has n connected components on the whole S'_{Δ} (S) as well.

Assumption 25 resembles of the weak unimodality condition.[9] For that case Assumption 25 says that if there is only one peak where the density reaches the level p on R, then there will be also only one peak on S'_Δ from the data for the same level p (the $g^1(c)$ is unimodal). Note that it does not say anything about the behavior of the density on the rest of S or outside of the identified regions. It is important to leave some space between the isodensity and the boundary of R. Without that condition we would not rule out the case that the repeating sets may be (periodic) parts of the same continuous line.

Remark.{ On the one hand, Assumption 25 is a generalization to Kline (2015), because it handles more than (weakly) unimodal distributions. On the other hand, Kline (2015) assumes that the mode of $\tilde{\epsilon}$ is in $S \cap S'_\Delta$, not on its subset $(R \cup G'_\Delta)$, so his assumption is less restrictive. However, that assumption is in general non-testable, exactly because it allows the mode to be in the region of the intersect that is not R or G'_Δ. }

Proposition 8 (Identification via matching levels.)

Under Assumption 17-23 and 25, Δ is identified.

The proof is included in Appendix A.

The following corollaries extend the result in some directions.

Corollary 4

[9] "Weak" unimodality means that there may be several local maxima of the density, but there is a unique global maximum. Strong unimodality means that there is a unique local maximum

If

1. *$\tilde{\epsilon}$ is weakly n-modal on S (S'_Δ) as well as on G'_Δ (R),*
2. *all the modes correspond to the same density value,*
3. *and Assumptions 17-23 are satisfied,*

then Δ is identified.

The same is of course not required in the case if one assumes there is one local maximum (strong modality), because the possibility of peaks with multiple values is ruled out as well.

Corollary 5

If

1. *$\tilde{\epsilon}$ is strongly n-modal on S (S'_Δ) as well as on G'_Δ (R),*
2. *and Assumptions 1-6 are satisfied,*

then Δ is identified.

Proof. The corollaries are direct consequences of Proposition 8. □

3.4.2 Restriction on the slope of the isodensity and the gradient

One might ask, what happens if there is no mode on the sets R and G'_Δ. In this section I give an example for another type of restriction that gives identification

if the modes are relatively far away from the two benchmark sets, and the ϵ-s are not too negatively dependent from each other. To be able to apply the convenient definitions of standard calculus, we assume for this section that the joint density of $\tilde{\epsilon}$ is smooth enough and that the isodensity sets are in fact curves.

Assumption 26 (Smoothness and variation of f_ϵ.)

The following is true for the density of ϵ:

1. *(smoothness) The joint and marginal probability density functions of the ϵ random vector are twice differentiable.*

2. *(no thick isodensity sets) For every c in S and S'_Δ for any open υ-ball ($\upsilon > 0$) around c there is a $c' \neq c$ point in the ball such that $f_{\tilde{\epsilon}}(c) \neq f_{\tilde{\epsilon}}(c')$.*

Choose a point Z from some I_p curve[10] on the set R (G'_Δ), and observe the slope of the isodensity curve at that point. If the isodensity corresponding to the same p value on the S'_Δ (S) defines a strictly convex upper-contour set (aside from the obviously weakly convex boundaries), then we can match the point Z at most with two points Z' and Z'' on S'_Δ (S). This is because a strictly convex set can only have two tangents with the same slope. Having strictly convex upper-contour sets also implies that if there are two tangents with the same slope, at one tangential point the gradient of the joint pdf will be negative, while at the other point it will be positive. This means that assuming strictly quasi-concave joint density gives a well-defined problem if the information is available whether the point is on the

[10] I_p should not consists of only one point.

increasing or the decreasing part of the density. This characteristic of point Z is well-defined and characterized by the next assumption and the first lemma.

Remark.{ The previous assumption is not testable, but the convexity of a given level set is a refutable assumption. Strict quasi-concavity of the pdf below is a much stronger condition than we need. For the identification argument to go through there is enough to have one strictly convex isodensity curve that is present on both sets. Moreover, the same argument works if we assume quasi-convexity, or a presence of a unique inflection point for at least one isodensity curve. In the latter case the inflexion point would play the role of the "landmark", similarly to the mode in the previous section. }

Intuitively, the strict convexity of the level sets of the joint probability density function depends on the shape of the marginals and the dependence structure of the two component. This intuition is captured by the copula representation of the joint pdf, which uniquely exists by Sklar's theorem and Assumption 26.

Assumption 27 (log-concavity)

The joint pdf admits the representation $f_\epsilon(c_1, c_2) = \gamma(c_1, c_2) \cdot f_1(c_1) f_2(c_2)$*, where the marginals* $f_i(c_i)$ *are strictly log-concave, while the copula-density* $\gamma(c_1, c_2)$ *is weakly log-concave.*

Proposition 9 (Identification via matching slopes.)

If Assumption 17-23 and 26-27 are satisfied, then Δ *is identified.*

After realizing that given the strict quasi-concavity of the joint pdf of the

unobservables gives identification, one only need to prove that log-concavity of the terms will imply that the product is quasi-concave. For this from Prekopa (1980) we know that the log-concave functions are quasi-concave, and that log-concavity is preserved by multiplication. The proof for the case with independence (when $\gamma(.,.) = 1$) is included in Appendix A.

Log-concavity is a widely-used assumption in economics (for a review see Bagnoli and Bergstrom 2005), and many of the well-known distributions are in this group (normal, extreme value, light-tailed Weibull, gamma, beta etc). The only very restrictive property of this group of distributions is that the density must have an exponentially vanishing tail. Heavy-tailed distributions, even the log-normal or the Pareto-distribution are not part of this family.

Remark.{ As in the previous remark, assuming for example independent log-convex distributions would work as well. If the researcher knows that on the R and the S'_Δ the density is strictly decreasing, then any two quasi-concave marginals will produce the required convex measure by Borell (1975). If one would like to add a dependence structure, restrictions on the copula pdf are needed along the same lines as in the previous case, using results related to uniform quasi-concavity Prekopa et al (2011). This would remedy the problem mentioned in the previous paragraph. }

3.5 Conclusion

In this paper we showed that with two unique potential outcomes the Nash-equilibrium assumption in a complete information game restricts the joint density of

the unobservables in such a way that the identification problem is described by the photo-stitching intuition. We concluded that the identification of the interaction terms is achieved with a much greater class of distributions that the literature has proven before, while maintaining that the observables are of bounded support. The framework developed here uses more information of the data and of the standard assumptions made in the literature, so it reveals more of the nature of the Nash-equilibrium assumption with unique outcomes. For future research the arguments used here open the possibility to define a more efficient estimator than the one introduced by Kline (2015) without relying on the large support assumption (see Appendix B). As another application of the framework presented above, more can be said about the case when the econometrician only have discrete observables at her disposal.

3.6 References

Bagnoli, M., and Bergstrom, T. (2005). Log-concave probability and its applications. Economic theory 26(2), 445-469.

Bajari, P., Hong, H., and Ryan, S. P. (2010). Identification and estimation of a discrete game of complete information. Econometrica 78(5), 1529-1568.

Berry, S., and Tamer, E. (2006). Identification in models of oligopoly entry. Econometric Society Monographs, 42, 46.

Berry, S., and Reiss, P. (2007). Empirical models of entry and market structure. Handbook of industrial organization, 3, 1845-1886.

Bjorn, P. A., and Vuong, Q. H. (1984). Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation. IDEAS RePec WP (No. 537).

Borell, C. (1975). Convex set functions in d-space. Periodica Mathematica Hungarica, 6(2), 111-136.

Bresnahan, T. F., and Reiss, P. C. (1991a). Entry and competition in concentrated markets. Journal of Political Economy, 977-1009.

Bresnahan, T. F., and Reiss, P. C. (1991b). Empirical models of discrete games. Journal of Econometrics 48(1), 57-81.

Brown, M., and Lowe, D. G. (2007). Automatic panoramic image stitching using invariant features. International journal of computer vision, 74(1), 59-73.

Ciliberto, F., and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. Econometrica, 77(6), 1791-1828.

de Paula, A., and Tang, X. (2012). Inference of signs of interaction effects in simultaneous games with incomplete information. Econometrica, 80(1), 143-172.

Fox, J. T., and Lazzati, N. (2015). Identification of Discrete Choice Models for Bundles and Binary Games. Working Paper.

Kline, B. (2015). The empirical content of games with bounded regressors. UT Austin Working Paper.

Mazzeo, M. J. (2002). Product choice and oligopoly market structure. RAND Journal of Economics, 221-242.

Manski, C. F. (1988). Identification of binary response models. Journal of the American Statistical Association, 83(403), 729-738.

Newey, W. K., and McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of econometrics, 4, 2111-2245.

Prekopa, A. (1980). Logarithmic concave measures and related topics. Stochastic programming. MAH Dempster, Academic Press, 63-82.

Prekopa, A., Yoda, K. and Subasi, M. M. (2011). Uniform quasi-concavity in probabilistic constrained stochastic programming. Operations Research Letters, 39(3), 188-192.

Singh, A., Scott, C., and Nowak R. (2009): Adaptive Hausdorff estimation of density level sets. The Annals of Statistics, 37(5B), 2760-2782.

Soetevent, A. R., and Kooreman, P. (2007). A discrete choice model with social interactions: with an application to high school teen behavior. Journal of Applied Econometrics, 22(3), 599-624.

Szeliski, R. (2006). Image alignment and stitching: A tutorial. Foundations and Trends in Computer Graphics and Vision, 2(1), 1-104.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. The Review of Economic Studies, 70(1), 147-165.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. The Annals of Statistics, 25(3), 948-969.

3.7 Appendix A

3.7.1 Proof of Proposition 7

More Notation

First we introduce some new notation that mathematically describes our version of the photo-stitching problem, and prove a claim on some newly defined objects that will be important for the proof of Proposition 7. This section is a detour to make the proof of identification and the discussion of the photo-stitching problem less obscure.

Denote the translation of S by the location vector $0 < \delta \leq d$ as $S'_\delta = S + \delta$, where $d = (d_1, d_2)$ are the dimensions of D, the maximum possible values for the interaction effects. The rectangle S'_δ is given by its meet and join, $\underline{c}'_\delta = \underline{c} + \delta$ and $\bar{c}'_\delta = \bar{c} + \delta$, respectively. If we superpose the meet of S and D, the largest rectangle with sides parallel to the axes in the upper-right corner of S that does not have more than one common point with D is called R, and it is characterized by its meet $(\underline{c} + d)$ and its join $(\bar{c})$. Similarly, if we superpose the join of D and S'_δ, the largest rectangle with sides parallel to the axes in the lower-left corner of S'_δ that does not have more than one common point with this D is called G'_δ, and it is characterized by its meet $(\underline{c}'_\delta)$ and its join $(\bar{c}'_\delta - d)$.

Figure 3.4 summarizes all this notation. The following lemma states an important characteristic of R and G'_δ that will be used in the proof of the main result.

Lemma 15

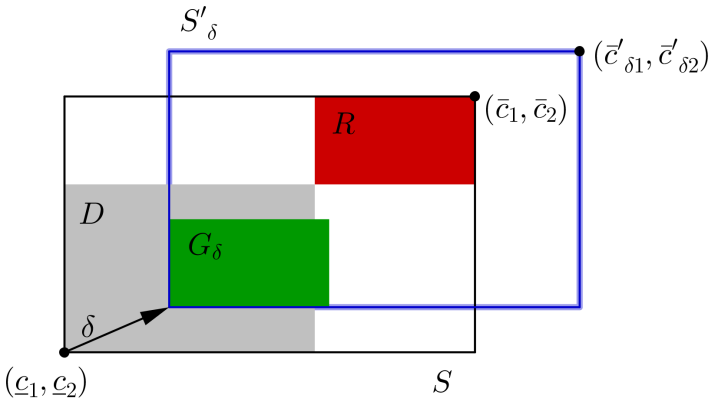


Figure 3.4: Notation summarized for the proofs. The red set is R, the green is G'_δ, while the set of the possible δ vectors is in the gray set D (starting from $\underline{c}$). The δ is the hypothetical vector of interaction effects by which S rectangle (black line) is translated into S'_δ (blue line). δ could point to any point within D, the gray set.

Under standing assumptions, R and G'_δ exists for all $\delta \in D$, and

$$R = \bigcap_{\delta \in D} S \cap S'_\delta,$$
$$G'_\delta = \bigcap_{\delta \in D} [S \cap S'_\delta - \delta] + \delta = G + \delta.$$

Where $S \cap S'_\delta - \delta$ is the translation of the intersect of S and S'_δ by the location vector $-\delta$. That is, R (G'_δ) is the largest part of S (S'_δ) such that R (the evaluation of G'_δ) is in the intersect of S and S'_δ for all admissible δ vectors.

Proof. By definition and Assumption 6, R and G'_δ always exist, and are of the same non-zero size for every δ. Similarly, since the δ lives in D, S and S'_δ always have an intersect that is also a non-zero measure rectangle. More importantly, for

any admissible $\delta \in D$

$$R \subseteq S \cap S'_\delta, \tag{3.10}$$

$$G'_\delta \subseteq S \cap S'_\delta. \tag{3.11}$$

For this to see one only need to consider that

$$\underline{c}'_\delta = \underline{c} + \delta \leq \underline{c} + d,$$
$$\bar{c}'_\delta - d = \bar{c} + \delta - d \leq \bar{c},$$

which is trivial, since by assumption $\delta \leq d$. Lemma 2 says that R is actually the supremum of the rectangles on S that are in the intersect of S and S'_δ for all δ, and G'_δ is the supremum of the rectangles on the second take, S'_δ that are subsets of the intersect os S and S'_δ for all δ. This means that indeed, these are the right definitions for benchmark sets, because they are the largest sets on the two "photos" that are on the overlapping region for all possible δ. To see this, assume that there is a point Q on S (resp. S'_δ) such that $Q \in S \cap S'_\delta$ for all δ, but $Q \notin R$ (or not in G'_d). However, while $\delta = d$ is an admissible vector, $G'_d = R = S \cap S'_d$ for this case, so Q cannot be in the intersect by assumption, which is a contradiction. □

Sufficiency of Assumption 24

The identification is based on the fact that both $P[(0,0)|c]$ and $P[(1,1)|c]$ give us the distribution of the same random variable on the intersect of S and S'_Δ. Since $S'_\Delta = S + \Delta$ is just a shift/translation of S by the small[11] vector Δ, the

[11] relative to S by Assumption 23

intersect must exist. Following the argument in the previous section, it is clear that since the Δ vector is positive, and it takes values in D with dimensions strictly smaller than the sides of S, the upper-right corner of S and the lower-left corner of S'_Δ will be a part of the intersect (the overlapping region). If one takes any $\delta \in D$ location vector, the implied intersect region is a product set of the following intervals

$$S \cap S'_\delta = [\underline{c}'_1, \underline{c}'_1 - \delta_1 + s_1] \times [\underline{c}'_2, \underline{c}'_2 - \delta_2 + s_2] = [\underline{c}_1 + \delta_1, \bar{c}_1] \times [\underline{c}_2 + \delta_2, \bar{c}_2],$$

which are the upper-right rectangle on S and the lower-left rectangles on S'_δ of the size $s_1 - \delta_1 \times s_2 - \delta_2$. If $\delta = \Delta$, then the predicted densities must be the same on these two rectangles. That is, the densities calculated by the $(0,0)$ conditional probabilities from the observations with c values around c must be the same as the density predicted by the $(1,1)$ conditional probabilities from the observations with c values around $c + \Delta$. Mathematically, using the definition of g^i_ϵ, the following must be true if $\delta = \Delta$ (the true value):

$$\sup_{c \in S \cap S'_\delta} |g^1_\epsilon(c - \delta) - g^0_\epsilon(c)| = 0. \tag{3.12}$$

Note that above we used that the Radon-Nikodym derivatives (or here simply the limits on $\mathbb{R}^2$) are unique for finite positive measures.

It follows that if there is a unique $\delta^* \in D$ that satisfies the criterion in equation (3.12), then $\delta^* = \Delta$, the true interaction effect. For identification, we need to rule out that there are two admissible δ-s for which (3.12) are true. This

condition would be that there is no $\delta^1 \neq \delta^2$ in D such that

$$\sup_{c \in S \cap S'_{\delta^1}} |g^1_{\tilde{\epsilon}}(c - \delta^1) - g^0_{\tilde{\epsilon}}(c)| + \sup_{c \in S \cap S'_{\delta^2}} |g^1_{\tilde{\epsilon}}(c - \delta^2) - g^0_{\tilde{\epsilon}}(c)| = 0.$$

The assumption that this condition would generate is less restrictive than Assumption 24, but it has even less interpretation and implies a non-standard estimation problem (the δ-s are included on the set of potential maximizing values) for future purposes. For this reason I will use a sufficient assumption that implies this.

We know from the previous section that the largest rectangle on S that is a subset of the intersect of S and S'_δ for all $\delta \in D$ is R. Similarly, the largest rectangle in the lower-left corner of S'_δ (as a function of $\underline{c}'$) that is in the intersect for all $\delta \in D$ is G'_δ. If we superimpose S'_δ with S, then the image of G'_δ will be G by definition. Since both the G'_Δ and R is a subset of $S \cap S'_\Delta$, the condition from equation (3.12) implies that if $\delta = \Delta$, then

$$\sup_{c \in (G'_\delta \cup R)} |g^1_{\tilde{\epsilon}}(c - \delta) - g^0_{\tilde{\epsilon}}(c)| = 0.$$

which is true if and only if

$$\sup_{c \in G'_\delta} |g^1_{\tilde{\epsilon}}(c - \delta) - g^0_{\tilde{\epsilon}}(c)| = 0$$
$$\sup_{c \in R} |g^1_{\tilde{\epsilon}}(c - \delta) - g^0_{\tilde{\epsilon}}(c)| = 0,$$

but then after plugging in $G = G'_\delta - \delta$ for the first equation to denote the same set of c-s we get that

$$\sup_{c \in G} |g^1_{\tilde{\epsilon}}(c) - g^0_{\tilde{\epsilon}}(c + \delta)| = 0$$
$$\sup_{c \in R} |g^1_{\tilde{\epsilon}}(c - \delta) - g^0_{\tilde{\epsilon}}(c)| = 0.$$

Assumption 24 rules out exactly the case when these two equations are satisfied, and note that every object in the conditions above are observed/identified. So Assumption 17-24 implies that there is a unique $\delta \in D$ that satisfies equation (3.12), which means we have identification. QED.

3.7.2 Proof of Corollary 3

The large support assumption is the limiting case when the area of S goes to infinity. Given that we hold the D constant, this would also mean that the area of R and G goes to infinity as well by definition. If Assumption 24 is not satisfied, then there are two vectors $\delta^1 \neq \delta^2$ for which every $c \in G$ and $c' \in G - \delta_1$ satisfies

$$g^0(c + \delta_1) = g^0(c + \delta_2) \Leftrightarrow g^0(c') = g^0(c' + \delta^2 - \delta^1),$$

for all c'.

Now assume that the ϵ's support (defined as the subset of $\mathbb{R}^2$ where it may be realized with non-zero probability) is unbounded.[12] This means that in the limit, the density on G would be perfectly periodic in the direction $\delta^2 - \delta^1$. But then it cannot be a proper density, because if the pdf would be integrable, it would integrate to ∞, unless its essential supremum (with respect to the Lebesgue-measure) is zero.

This is true because of the following argument. If the density integrates to a positive probability on some finite subset[13] $B_G \in G + \delta_1$, then it will integrate to

[12] If the support is bounded, as S tends to the whole plane, the problem is trivial, because the boundary point of the support is identified.

[13] I further apply the notational convention that if there is a set $A \in \mathbb{R}^2$ and a (position) vector v at the same space, then $A + v = \{a \in \mathbb{R}^2 : a - v \in A\}$.

the same positive number for the set $B_G + n(\delta^2 - \delta^1)$ for every n. And so *in the limiting case,* since $\delta^2 - \delta^1$ is not a null-vector by assumption, for high enough k the intersect $B_G \cap B_G + kn(\delta^2 - \delta^1)$ will be the empty set for all n, and

$$\mu(G) \geq \mu\left[\lim_{N\to\infty}\sum_{n=0}^{N}(B_G + kn(\delta^2 - \delta^1))\right] = \lim_{N\to\infty}\sum_{n=0}^{N}\mu(B_G + kn(\delta^2 - \delta^1)) =$$
$$= \lim_{N\to\infty} N\mu(B_G) = \infty.$$

This means we cannot have any (non-zero area) subset of G to have a positive measure. But then since ϵ has a proper density, the $\tilde{\epsilon}$-measure of G would be zero, and by the same argument the measure of R as well, which would contradict to the assumption that ϵ has unbounded support.

3.7.3 Proof of Proposition 8

I will only prove that Assumption 25 implies 24 in relation to R and S'_Δ, as the other case is analogous. Assume that Assumption 24 is violated. Then there are two vectors $\delta^1 \neq \delta^2$ in D for which there are two congruent sets $R'_1 = R'_2 - \delta^2 + \delta^1$ in S'_δ with the exact same isodensity curve I_p as on R. If there is such p, because of the assumption that the I_p on R's interior would be covered by a closed set it must be true that the non-connected connected components of that set remain disconnected from each other and from any other component of I_p outside of the images of R when translated to S'_Δ. So all that is remained to see that after the translation of R on $\mathbb{R}^2$ by the two vectors, the number of connected components of I_p on the union of R'_1 and R'_2 cannot be smaller than $n + 1$.

For this, consider the following argument. If there are $R'_1 \neq R'_2$ such that

the connected components of I_p on each set are strict subsets of the interior, then $(R_1' \backslash R_2', R_2')$ is a partition of $R_1' \cup R_2'$ (the two sets are disjoint). First, $R_1' \backslash R_2'$ must have at least one (nonempty) connected component of I_p, else the $R_1' = R_2'$, because the vector of translation is given by one point and its image. Second, because of the additional assumption concerning the boundary, we know that the connected components on R_2' are disconnected from the connected components of I_p on $R_1' \backslash R_2'$, so on the union of the two sets I_p must have at least $n + 1$ connected components.

After making a symmetric argument for G'_Δ and S, we can conclude that if Assumption 7 is violated, Assumption 8 cannot hold either. This means that Assumption 25 implies Assumption 24.

3.7.4 Proof of Proposition 9

By Prekopa (1980) we know that the log-concave functions are quasi-concave, and that log-concavity is preserved by multiplication. This means that if the marginals and the copula pdf are strictly log-concave, then the joint pdf will be strictly quasi-concave. So by the argument above, Assumption 24 is satisfied, and identification of Δ is achieved. I only include here the proof of log-concavity in the case of independence, adding the copula is analogous to the argument below.

In this case we know

$$f_{\tilde{\epsilon}}(c_1, c_2) = f_1(c_1) f_2(c_2),$$

and strict log-concavity means that for $0 < \lambda < 1$

$$f_i(\lambda c_i + (1 - \lambda) c_i') < f_i(c_i)^\lambda f_i(c_i')^{1-\lambda}.$$

If we put these together

$$f_{\tilde{\epsilon}}(\lambda c_1 + (1-\lambda)c_1', \lambda c_2 + (1-\lambda)c_2') = \prod_i f_i(\lambda c_i + (1-\lambda)c_i') <$$
$$< \left[\prod_i f_i(c_i)\right]^{\lambda} \left[\prod_i f_i(c_i')\right]^{1-\lambda} =$$
$$= f_{\tilde{\epsilon}}^{\lambda}(c_1, c_2) f_{\tilde{\epsilon}}^{1-\lambda}(c_1', c_2'),$$

which is the sufficient and necessary condition for the strict (log-)concavity of a continuous function[14].

[14]It is enough to show for $\lambda = 0.5$ in fact.

3.8 Appendix B

Due to lack of time and space, the following results are preliminary, and the arguments are heuristic.

3.8.1 Corresponding estimator of Δ

Define the estimator $\hat{\Delta}$ as

$$\hat{\Delta} = \arg\min_{\delta \in D} \left\{ \sum_{B_n^R \in R} \left(P^1(B_n^R) - P^0(B_n^R) \right)^2 + \sum_{B_n^G \in G} \left(P^1(B_n^G) - P^0(B_n^G) \right)^2 \right\}, \tag{3.13}$$

where $B_n^{R,G}$ are disjoint rectangles in R and G respectively, such that $\cup B_n^R = R$, $\cup B_n^G = G$, and $B_n^{R,G} = [\underline{c}_1^{B^{R,G}}, \bar{c}_1^{B^{R,G}}] \times [\underline{c}_2^{B^{R,G}}, \bar{c}_2^{B^{R,G}}]$. That is, for every n the R and G are partitioned into product cylinders (rectangles). The quantities in the

objective function are calculated as follows:

$$P^0(B_n^R) = \hat{P}\left[(0,0)|c_1 = \bar{c}_1^{B_n^R}, c_2 = \bar{c}_2^{B_n^R}\right] - \hat{P}\left[(0,0)|c_1 = \bar{c}_1^{B_n^R}, c_2 = \underline{c}_2^{B_n^R}\right] - \quad (3.14)$$
$$- \hat{P}\left[(0,0)|c_1 = \underline{c}_1^{B_n^R}, c_2 = \bar{c}_2^{B_n^R}\right] + \hat{P}\left[(0,0)|c_1 = \underline{c}_1^{B_n^R}, c_2 = \underline{c}_2^{B_n^R}\right],$$

$$P^1(B_n^R) = \hat{P}\left[(1,1)|c_1 = \bar{c}_1^{B_n^R} - \delta_1, c_2 = \bar{c}_2^{B_n^R} - \delta_2\right] - \quad (3.15)$$
$$- \hat{P}\left[(1,1)|c_1 = \bar{c}_1^{B_n^R} - \delta_1, c_2 = \underline{c}_2^{B_n^R} - \delta_2\right] -$$
$$- \hat{P}\left[(1,1)|c_1 = \underline{c}_1^{B_n^R} - \delta_1, c_2 = \bar{c}_2^{B_n^R} - \delta_2\right] +$$
$$+ \hat{P}\left[(1,1)|c_1 = \underline{c}_1^{B_n^R} - \delta_1, c_2 = \underline{c}_2^{B_n^R} - \delta_2\right],$$

$$P^1(B_n^G) = \hat{P}\left[(1,1)|c_1 = \bar{c}_1^{B_n^G}, c_2 = \bar{c}_2^{B_n^G}\right] - \quad (3.16)$$
$$- \hat{P}\left[(1,1)|c_1 = \bar{c}_1^{B_n^G}, c_2 = \underline{c}_2^{B_n^G}\right] -$$
$$- \hat{P}\left[(1,1)|c_1 = \underline{c}_1^{B_n^G}, c_2 = \bar{c}_2^{B_n^G}\right] +$$
$$+ \hat{P}\left[(1,1)|c_1 = \underline{c}_1^{B_n^G}, c_2 = \underline{c}_2^{B_n^G}\right],$$

$$P^0(B_n^G) = \hat{P}\left[(0,0)|c_1 = \bar{c}_1^{B_n^G} + \delta_1, c_2 = \bar{c}_2^{B_n^G} + \delta_2\right] - \quad (3.17)$$
$$- \hat{P}\left[(0,0)|c_1 = \bar{c}_1^{B_n^G} + \delta_1, c_2 = \underline{c}_2^{B_n^G} + \delta_2\right] -$$
$$- \hat{P}\left[(0,0)|c_1 = \underline{c}_1^{B_n^G} + \delta_1, c_2 = \bar{c}_2^{B_n^G} + \delta_2\right] +$$
$$+ \hat{P}\left[(0,0)|c_1 = \underline{c}_1^{B_n^G} + \delta_1, c_2 = \underline{c}_2^{B_n^G} + \delta_2\right].$$

The number of rectangles in the partitions (h(n)) tends to infinity with the number of observations, but a smaller rate, we shall have $h(n) = o(\sqrt{n})$. Lastly, when there are n observations, the conditional probabilities of the respected outcomes are

$$\hat{P}\left[(a,b)|c_1 = \tilde{c}_1, c_2 = \tilde{c}_2\right] = \frac{\sum_{i=1}^n \mathbf{1}_{[s_1=a,s_2=b]}\mathbf{1}_{[c_1=\tilde{c}_1,c_2=\tilde{c}_2]}}{\sum_{i=1}^n \mathbf{1}_{[c_1=\tilde{c}_1,c_2=\tilde{c}_2]}}.$$

3.8.2 Identification of the slope parameters

The identification of the slope parameters does not require the differentiation of the cdf/survival function. The intuition why the equilibrium-outcome probabilities pin down the slope parameters (up to a scale) comes from the meaning of exogeneity and the assumption that the $F_{\tilde{\epsilon}}$ is strictly increasing. As a normalization, we assume that the coefficient on the first observable in both player's payoff-function is 1. As Kline (2015) also discusses, to be able to do this normalization and proceed with the arguments in the main part of the paper, a sign-restriction and the assumption that the coefficient is not zero is necessary. All in all, it is enough to assume that the researcher knows the sign of one observable, a condition that is typically fulfilled in our example, taking market size as such variable for instance. If this is true, one can always redefine the given observable as its negative inverse, and then do the normalization.

After assuming $F_{\tilde{\epsilon}}$ is strictly increasing in both direction on S (the density of ϵ is bounded away from zero a.e.), from equation (3.3) we have that

$$P[(0,0)|c^1] = P[(0,0)|c^2] \Leftrightarrow F_{\tilde{\epsilon}}(c^1) = F_{\tilde{\epsilon}}(c^2) \Leftrightarrow c^1 = c^2.$$

The information the observables give with respect to the payoff of player i is fully incorporated into the corresponding $c_i(x_i)$ value, and so by exogeneity, the combination of x_i values that amount to the same $c_i(x_i)$ will give the same probabilities. After assuming strictly increasing $F_{\tilde{\epsilon}}$ (on the relevant set, S), the relationship becomes valid for both directions. A rather technical condition we need in addition is that the jth observable ($x_i[j]$) is not "too discrete" and/or the corresponding

slope parameter ($\beta_i[j]$) is small enough in magnitude, so that the variable can take at least two values such that the two corresponding $c_i(x_i)$ values are still on S (they are observed). This caveat does not present itself with fully continuous set of regressors, of course. Either way, identification comes from the fact that after restricting the observations (games) for a given value of $P[(0,0)] = p^0$ and a particular set of x_2, for these observations

$$\beta_1 x_1 - c_1^0 = 0,$$

for some constant c_1^0. Or after rearranging,

$$\beta_1^* x_1^* = x_1[1], \tag{3.18}$$

where the β_1^* is the vector of slope parameters after trimmed from the first entry (that was a "1") and augmented with $-c_1^0$ instead, and x_1^* is a matrix with m_1 columns, in which the first column is a vector of ones and the last $m_1 - 1$ columns are the values of the observables except for the first regressor (which was taken to the other side of the equation). Note that the equation (3.18) leads to a structure similar to the OLS regression, so using the classical results, after assuming the $E[(x_1^*)' x_1^*]$ is invertible, β_1^* is identified. The above matrix may be singular for this particular probability value p^0, but cannot be singular for all of them if we assume the usual non-singularity condition, the second part of Assumption 5.[15] After the identification of the slope parameters for Player 1, the β_2 vector's case is symmetric.

[15] We need a slightly modified form that includes the constant as well, to be exact.

This identification strategy is almost exactly the same as in Manski (1988) the identification of the binary choice models for the strong exogeneity case. The probit-tobit parallel also suggests that once the conditional expected values of the payoffs are observed (or at least the differences of them - e.g. in a lottery setting), then the coefficients will be identified (not only up to scale).

Another intuition for $m_1 = 2$ is to plot the isoprobability sets (the equivalence classes of points on the observed $(x_1[1], x_1[2])$ space where the probabilities of getting $(0, 0)$ outcomes equal) while holding the observables of the other player constant. Those curves must be lines by our assumptions, and the slope of the lines must be $-\beta_1[2]$ (second entry of the vector of slope coefficients).

Bibliography

Abrevaya, J. and Shin, Y. (2011). Rank estimation of partially linear index models. The Econometrics

Abrevaya, J. (2000). Rank estimation of a generalized fixed-effects regression model, Journal of Econometrics, Volume 95, Issue 1, 2000, pp. 1-23.

Abrevaya, J. (1999). Rank estimation of a transformation model with observed truncation. The Econometrics Journal, 2: 292-305.

Ahn, H., (1995). Nonparametric two-stage estimation of conditional choice probabilities in a binary choice model under uncertainty, Journal of Econometrics, Volume 67, Issue 2, 1995, 337-378,

Ahn, H., Ichimura, H., Powell, J. L. and Ruud, P. A. (2015). Simple Estimators for Invertible Index Models, WP

Ahn, H. and Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. Journal of Econometrics, Volume 58, Issues 1–2, 1993, pp. 3-29.

Albert, R., Barabasi, A. L. (2002). Statistical mechanics of complex networks. Reviews of modern physics, 74(1), 47.

Andrews, D. W.K. (1989). Asymptotics for Semiparametric Econometric Models: I. Estimation, Cowles Foundation Discussion Papers 908R, Cowles Foundation for Research in Economics, Yale University, revised Aug 1990.

Andrews, D. W. K. (1994a). Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity, Econometrica, Econometric Society, vol. 62(1), pages 43-72, January.

Andrews, D.W.K., (1994b). Nonparametric Kernel Estimation for Semiparametric Models, Econometric Theory, Cambridge University Press, vol. 11(03), pages 560-586, 1995 June.

Arcones, M. A., Gine, E. (1993,1991). Limit Theorems for U-Processes. Ann. Probab. 21 (1993), no. 3, 1494-1542.

Audibert, JY., Tsybakov, B. (2007). Fast learning rates for plug-in classifiers. Ann. Statist. 35 (2007), no. 2, 608-633.

Bagnoli, M., and Bergstrom, T. (2005). Log-concave probability and its applications. Economic theory 26(2), 445-469.

Bajari, P., Hong, H., and Ryan, S. P. (2010). Identification and estimation of a discrete game of complete information. Econometrica 78(5), 1529-1568.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. Journal of Econometrics, 2015, vol. 186, issue 2, 345-366

Berry, S., and Tamer, E. (2006). Identification in models of oligopoly entry. Econometric Society Monographs, 42, 46.

Berry, S., and Reiss, P. (2007). Empirical models of entry and market structure. Handbook of industrial organization, 3, 1845-1886.

Bjorn, P. A., and Vuong, Q. H. (1984). Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation. IDEAS RePec WP (No. 537).

Blundell, R. W. and Powell, J.L. (2004). Endogeneity in Semiparametric Binary Response Models. Restud Review of Economic Studies (2004) 71, 655–679

Borell, C. (1975). Convex set functions in d-space. Periodica Mathematica Hungarica, 6(2), 111-136.

Bresnahan, T. F., and Reiss, P. C. (1991a). Entry and competition in concentrated markets. Journal of Political Economy, 977-1009.

Bresnahan, T. F., and Reiss, P. C. (1991b). Empirical models of discrete games. Journal of Econometrics 48(1), 57-81.

Brown, M., and Lowe, D. G. (2007). Automatic panoramic image stitching using invariant features. International journal of computer vision, 74(1), 59-73.

Cavanagh, C. and Sherman, R. P. (1998). Rank estimators for monotonic index models. Journal of Econometrics, 84(2):351–381.

Charbonneau, K. B. (2017), Multiple fixed effects in binary response panel data models. The Econometrics Journal, 20: S1-S13.

Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding exponential random graph models. Annals of Statistics, 41(5): 2428–2461.

Ciliberto, F., and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. Econometrica, 77(6), 1791-1828.

de Paula, A., and Tang, X. (2012). Inference of signs of interaction effects in simultaneous games with incomplete information. Econometrica, 80(1), 143-172.

de Paula, A. (2016). Econometrics of network models. Technical Report CWP06/16, CEMMAP.

de Paula, A., Richards-Shubik, S., and Tamer, E. (2015). Identification of preferences in network formation games. Technical Report CWP29/15, CEMMAP.

Dzemski, A. (2014). An empirical model of dyadic link formation in a network with unobserved heterogeneity". Technical report, University of Gothenburg.

Dragomir, S. S. (2015). Reverses of Schwarz inequality in inner product spaces with applications. Math. Nachr., 288: 730-742. doi:10.1002/mana.201300100

Drukker, D. M. and Stinchcombe, M. B. (2014). Regression efficacy and the curse of dimensionality. WP

Fan, Y., Han, F., Li, W., Zhou, X-W. (2017). On Rank Estimators in Increasing Dimensions, WP

Fox, J. T., and Lazzati, N. (2015). Identification of Discrete Choice Models for Bundles and Binary Games. Working Paper.

Froelich, M. (2006). Non-parametric regression for binary dependent variables. The Econometrics Journal, Vol. 9, No. 3 (2006), pp. 511-540

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. Journal of Business and Economic Statistics, 31(3):253 – 264.

Graham, B. S. (2015). An econometric model of link formation with degree heterogeneity. Technical Report 20341, National Bureau of Economic Research. Published in Econometrica (2017).

Graham, B. S. (2015). Methods of identification in social networks. Annual Review of Economics, 7:465 – 485.

Graham, B. S. (2016). Homophily and transitivity in dynamic network formation (No. w22186). National Bureau of Economic Research.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. Journal of Econometrics, 35(2-3), 303-316.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. Econometric Theory, 24(03), 726-748.

He, X. and Shao, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. The Annals of Statistics, 24(6):2608–2630.

Heckman, J. J. (2008). Econometric causality. International statistical review, 76(1), 1-27.

Hoderlein, S. and White, H. (2012,2010) Nonparametric identification in nonseparable panel data models with generalized fixed effects, Journal of Econometrics, Volume 168, Issue 2, 2012, Pages 300-314,

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics, 19(3):293 – 325.

Honore, B. E. and Powell, J. (2005). Pairwise difference estimators for nonlinear models. In Andrews, D.W.K., Stock, J.H. (Eds.) Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg, pages 520–553. Cambridge University Press.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 221–233. Berkeley, CA.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. The Annals of Statistics, 1(5):799–821.

Jackson, M. O. (2008). Social and Economic Networks. Princeton University Press.

Jackson, M. O. and Wolinsky, A. (1996). A strategic model of social and economic networks. Journal of Economic Theory, 71(1):44 – 74.

Jochmans, K. (2017): Semiparametric Analysis of Network Formation,Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2017.1286242

Jochmans, K. and Weidner M. (2017): Fixed-Effect Regressions On Network Data, Working Paper

Khan, S. and Tamer, E. (2007). Partial rank estimation of duration models with general forms of censoring. Journal of Econometrics, 136(1):251–280.

Kiefer, J. (1967). On Bahadur's representation of sample quantiles. The Annals of Mathematical Statistics, 38(5):1323–1342.

Kline, B. (2015). Identification of complete information games. Journal of Econometrics, 189(1), 117-131. Chicago

Kline, B. (2015). The empirical content of games with bounded regressors. UT Austin Working Paper.

Lovasz, L. (2012). *Large Networks and Graph Limits.* AMS, Colloquium Publications, Vol 60.

Liu, N., Xu, H. (2012). Semiparametric analysis of social interactions with homophily. UT Austin Working Paper.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. Journal of Econometrics, 3(3):205 – 228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. Journal of Econometrics, 27(3):313 – 333.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. Econometrica, 55(2):357 – 362.

Matzkin, Rosa. (1989). A Nonparametric Maximum Rank Correlation Estimator. Cowles Foundation WP.

Matzkin, R., (2007). Nonparametric identification. In: Handbook of Econometrics, 2007, vol. 6B, Chapter 73, Elsevier

Mazzeo, M. J. (2002). Product choice and oligopoly market structure. RAND Journal of Economics, 221-242.

Mele, A. (2015). A structural model of segregation in social networks. Technical report, John Hopkins University.

Menzel, K. (2015). Strategic network formation with many agents. Technical report, New York University.

Newey, W. (1997). Convergence rates and asymptotic normality for series estimators. Journal of Econometrics, 1997, vol. 79, issue 1, 147-168

Newey, W. K., and McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of econometrics, 4, 2111-2245.

Nolan, D., and Pollard, D., (1987). U-Processes: Rates of Convergence. The Annals of Statistics, Vol. 15, No. 2 (June, 1987), 780-799.

Pakes, A., and Pollard, D. (1989). Simulation and the Asymptotics of Optimization Estimators. Econometrica, 57(5), 1027-1057.

Prekopa, A. (1980). Logarithmic concave measures and related topics. Stochastic programming. MAH Dempster, Academic Press, 63-82.

Prekopa, A., Yoda, K. and Subasi, M. M. (2011). Uniform quasi-concavity in probabilistic constrained stochastic programming. Operations Research Letters, 39(3), 188-192.

Rudelson, M. (1999). Random Vectors in the Isotropic Position. Journal of Functional Analysis, Volume 164, Issue 1, 60-72,

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. John Wiley and Sons.

Shi, X., and Shum, M. (2017). Estimating Semi-parametric Panel Multinomial Choice Models using Cyclic Monotonicity, WP

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. Econometrica 61 123–137.

Sherman, R. P. (1994). Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators. Ann. Statist. 22

Sheng, S. (2014). A structural econometric analysis of network formation games. Technical report, UCLA.

Singh, A., Scott, C., and Nowak R. (2009): Adaptive Hausdorff estimation of density level sets. The Annals of Statistics, 37(5B), 2760-2782.

Snijders, T. A. B. (2011). Statistical models for social networks. Annual Review of Sociology, 37:131 – 153.

Soetevent, A. R., and Kooreman, P. (2007). A discrete choice model with social interactions: with an application to high school teen behavior. Journal of Applied Econometrics, 22(3), 599-624.

Stone, Ch. J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. Ann. Statist. 10 (1982), no. 4, 1040-1053.

Szeliski, R. (2006). Image alignment and stitching: A tutorial. Foundations and Trends in Computer Graphics and Vision, 2(1), 1-104.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. The Review of Economic Studies, 70(1), 147-165.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. The Annals of Statistics, 25(3), 948-969.

van der Vaart, A. and Wellner, J. (1996). Weak Convergence and Empirical Processes. Springer.

Wang, H. (2007). A note on iterative marginal optimization: a simple algorithm for maximum rank correlation estimation. Computational Statistics and Data Analysis, 51(6):2803–2812.